2002

Essays on the Bayesian inequality restricted estimation

Asli K. Ogunc
Louisiana State University and Agricultural and Mechanical College

ESSAYS ON THE BAYESIAN INEQUALITY RESTRICTED ESTIMATION

A Dissertation
Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in
The Department of Economics

by
Asli K. Ogunc
B.B.A., Marmara University, 1990
M.B.A., Western Michigan University, 1992
M.S., Louisiana State University, 1996
May 2002

DEDICATION

This work is dedicated to my father, Erden Kavaklioglu, who continues to strengthen, inspire, watch over and amaze me every moment of my life.

ACKNOWLEGEMENTS

I would like to thank all my family and friends who have supported me to in my pursuit for the doctorate in degree in economics. I am grateful to Drs. R. Carter Hill and M. Dek Terrell for exposing me to a *brave new world* and giving me an opportunity to dwell in it. I want to thank other committee members for their insights: Dr. W. Douglas McMillin, Dr. Lynn R. LaMotte, and Dr. Bogdan S. Oporowski.

I also would like to thank all my friends in the Economics Department of Louisiana State University, especially Dr. Janet Daniel, Mary Jo Neathery, Dr. Cagla Okten, Vera Tabakova, and Dan Teodorescu, who facilitated my long distance quest.

Finally, this dissertation would be impossible to complete without the sacrifices of my mother, Nur Kavaklioglu, support of my husband, Kurtay Ogunc, the inspiration of my father, Erden Kavaklioglu and my daughter, Patara Ögunc.

TABLE OF CONTENTS

ABSTRACT

Bayesian estimation has gained ground after Markov Chain Monte Carlo process made it possible to sample from exact posterior distributions. This research aims at contributing to the ongoing debate about the relative virtues of the Frequentist and Bayesian theories by concentrating on the qualitative dependent variable models. Two Markov Chain Monte Carlo (MCMC) methods have been used throughout this dissertation to facilitate Bayesian estimation, namely Gibbs (1984) sampling and the Metropolis (1953, 1970) Algorithm.

In this research, several Monte Carlo experiments have been carried out to better understand the finite sample properties of Bayesian estimator and its relative performance to Maximum Likelihood Estimation (MLE) in probit and poisson models. In addition, the performance of the estimators is compared when inequality restrictions are imposed on the coefficients of the models. The restrictions are imposed within the context of a Monte Carlo experiment for the probit model and applied to the real data in the poisson regression framework. The research demonstrates the ease with which the inequality restrictions on the coefficients of the probit and poisson models via the Gibbs sampler and Metropolis Algorithms, respectively.

It has been shown throughout the research that sample size has the largest impact on the risk of the parameters in both techniques. Bayesian estimation is very sensitive to prior specification even in the case of non-informative priors. Lowering the variance of the non-informative prior improves the Bayesian estimation, without significantly changing the nature of the distribution. In the cases where Bayesian prior variance is very large, MLE dominates the Bayesian in the almost all of the experimental designs. Whereas, when the prior variance is lowered, the improvement in the estimation process is remarkable.

In the constrained cases, the Bayesian estimator has lower variance and lower MSE when the restrictions are correct. As the specification error increases, the Bayesian estimator suffers more than the MLE. The increase in bias is more than the efficiency gain for the Bayesian case. The effects of changes such as the changes in the distribution of regressors, parameter values, collinearity, and their interactions warrant more investigation.

CHAPTER 1

BAYESIAN ESTIMATION

1.1 Introduction

The main contribution of this dissertation is to provide comparisons of Bayesian and Maximum Likelihood (ML) estimation techniques with and without inequality constraints. The Bayesian approach seeks to optimally combine information from two sources:

(i) information contained in the data in the form of a likelihood function, and

(ii) knowledge (objective) that is known from a theory/postulate or opinion (subjective) formed at the beginning of the research in the form of a prior.

The major opposition from non-Bayesian researchers is the subjectivity embedded in the prior. The theoretical foundations of economics as a scientific field form the basis for the description of priors within the context of this dissertation. As Leamer (1985) explains in the description of his methodology, I am also concerned with the sensitivity of inferences (in posterior means) to variations in assumptions (in the prior specifications of the parameter space). To this end, I extensively utilize sensitivity analysis with the hope that the Bayes Rule would provide a flexible means for using uncertain prior knowledge and combining disparate evidence. The use of Markov Chain Monte Carlo (MCMC) techniques allows us to apply Bayesian estimation to the problem of incorporating truncated priors.

Freedman (1986) makes the following statement as a non-Bayesian: “When drawing inferences from data, even the most hardbitten objectivist has to introduce assumptions and use prior information. The serious question is how to integrate that

information into the inferential process and how to test the assumptions underlying the analysis." The Bayesian methodology, in general, offers a solution to the problem posed by Freedman by incorporating prior information in a formal manner to generate complete posterior densities for parameters and predictive densities for future observations. Moreover, unlike non-Bayesian approaches, Bayesian methodologies that associate probabilities with hypotheses, models and propositions can deal with these important problems in an operational, reproducible manner.

Bayesian epistemology has two main elements as its formal apparatus:

(i) the use of the laws of probability as coherence constraints on rational degrees of belief (or degrees of confidence), and

(ii) the introduction of a rule of probabilistic inference, a rule or principle of conditionalization.

In general, what makes the Bayesian thinking unique is the conviction that an important principle governing rational changes in degrees of belief is the notion of conditionalizing in a generalized setting.[1] Jaynes (1985) provides a strong explanation as to what the Bayesian methodology intends to accomplish in the field of statistical estimation: "In Bayesian parameter estimation, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it. The width of the distribution is not intended to indicate the range of variability of the true values of the parameter. To the contrary, it indicates the range of values that are consistent with our prior information and data, and which honesty therefore compels us to admit as possible values. What is 'distributed' is not the

[1] Berger (1986) postulates that any frequentist answer is not inherently sensible for it lacks some plausible relationship to a meaningful conditional measure.

parameter, but the probability." In other words, Bayesians treat the parameter(s) of a given model as a random variable in the sense that one can assign to it a subjective probability distribution that describes his uncertainty about the actual value of the parameter.

Savage (1954) extended de Finetti (1937) on the notion of subjective Bayesianism to incorporate prior opinions, not prior knowledge, into scientific inference as a normative model. As opposed to objective Bayesians such as Rosenkrantz (1981), subjective Bayesians do not believe that rationality alone places enough constraints on one's prior probabilities to make them objective. As Jaynes puts it directly in the context of human intelligence: "Our brains are in possession of more principles than the robot's for converting raw information, semiquantitatively, into something which the computer can use."

The Bayesian approach summarizes information about the unknown parameter(s) in terms of a probability density function. The treatment of unknown parameters as if they were random variables provides a feedback mechanism to update our original beliefs about the parameter(s). The posterior distribution of the parameter(s) represents our revised belief and is calculated by combining data and prior knowledge. Mathematically, a Bayesian estimate of an unknown parameter is derived as the value that minimizes the posterior expected loss function. With this approach, the Bayesian estimate depends on the selected loss function as well as the prior distribution. For example, the Bayesian estimate under a quadratic loss function is the mean of the posterior distribution, and with a constant prior, the posterior distribution will be proportional to the likelihood function.

Until recently the extensive computational costs of evaluating the posterior distributions have hampered Bayesian econometrics. The emergence of MCMC techniques in the last ten years has led to the wider applications of the Bayesian methodology by facilitating the posterior calculations. The contribution of the current treatment of the Bayesian methodology to the growing literature will be in the form of imposing inequality restrictions on the parameters and measuring the estimation risk to compare the ML and Bayesian alternatives. The behavior of risk factors is empirically examined in response to changes in the design of the experiment in a Monte Carlo setting. To this end, the inequality constraints imposed on the parameter space provide an opportunity to utilize prior knowledge for making better inferences on the problem at hand[2]. Within the current setting, the methodology adheres to the Jeffreys-Wrinch Simplicity Postulate, which says that the simpler laws should have the greater prior probabilities (Jeffreys, 1967).

1.2 Bayes Theorem

The cornerstone of Bayesian methodology is the Bayes theorem[3], which is known as the principle of inverse probability. It helps us make probability statements about parameters after the sample has been taken. The conditional distribution of the parameters after observing the data is the posterior distribution that summarizes the prior

[2] Dorfman and McIntosh (1999) provide a theoretical result for Bayesian estimation subject to inequality constraints in the form of a lemma: "… Thus, the inequality restricted posterior mean will have a smaller second moment than the (restricted or unrestricted) maximum likelihood estimator." This is a direct result of the Bayesian approach giving zero weight to those regions of the parameter space that violate the restrictions. Refer to the paper for a proof of the lemma.

[3] Bayesian thinking offers a rationalistic theory of personal beliefs in the context of uncertainty and characterizes how an individual should act to avoid certain kinds of undesirable behavioral inconsistencies. In this sense, it yields a prescriptive (normative) proposition regarding modeling real phenomenon.

and the sample information. Let q denote a vector of parameters and *y* denote a vector of sample observations. The first step in deriving the posterior distribution is to calculate the joint probability distribution of the data and the parameters. This joint probability statement can be written as the product of the likelihood function, $f(y|q)$ and the joint prior distribution of the parameters, $P(q)$.

$$P(q, y) = f(y|q)P(q) \tag{1.1}$$

Using the basic property of the conditional distributions it is also true that,

$$P(q|y) = \frac{P(q, y)}{P(y)} \tag{1.2}$$

Substituting (1.1) into (1.2) will yield,

$$P(q|y) = \frac{f(y|q)P(q)}{P(y)} \tag{1.3}$$

The marginal distribution of the data can be regarded as a constant with respect to θ and therefore can be dropped in the calculation of the posterior, and instead a non-normalized posterior density can be used where $\propto$ denotes 'proportional to':

$$P(q|y) \propto f(y|q)P(q) \tag{1.4}$$

In other words, (1.4) could be phrased as:

Posterior information is proportional to sample information times prior knowledge

Attainment of the posterior is only the beginning of the research methodology since the statistical inference will be based on the posterior and predictive distributions that are derived using the Bayes' rule. However, to obtain the posterior we do need the data and the prior distribution. The choice of the prior distribution depends on the knowledge of the investigator as well as his willingness to incorporate beliefs and

theoretical postulates into the methodology. There are explicit rules for selecting prior distributions whether an informative or non-informative prior is preferred. In the spirit of the Simplicity Postulate, it is reasonable to begin with a simple case, a regression model with a constant term and a regressor. This model can be written as

$$y = \mathrm{X}b + e \tag{1.5}$$

where e is a $(T \times 1)$ vector of independent normally distributed random variables with zero mean and constant variance s^2, y is a $(T \times 1)$ vector of observations and X is $(T \times 2)$ matrix of observations on explanatory variables. s^2 and b_i are unknown parameters of this model and $q' = (b_i^{`}, s^2)$.

1.2.1 Analysis with Non-informative Priors

A non-informative prior is assigned to the parameters if the investigator does not have information on the parameter or does not want to use the prior information. In the case where s is known, only a prior for b is needed. The most common non-informative prior for the general linear model is,

$$P(b) \propto constant \tag{1.6}$$

Since we are assuming ignorance about the values of the unknown parameters, the posterior function in this case will be proportional to the likelihood function and will have the form

$$P(b|x) \propto f(x|b)P(b) \Rightarrow P(b|x) \propto f(x|b)$$

As a result the posterior distribution of b will be,

$$P(b|x) \sim N(b, s^2(\mathrm{X}'\mathrm{X})^{-1}) \tag{1.7}$$

The result in (1.7) is very similar to the form of its sampling theory counterpart, but is different in interpretation since the Bayesian approach treats the unknown parameters as if they were random, rather than the estimator. In the Classical framework, the estimators are evaluated by their performance in repeated samples, and probability is defined as the limiting frequency. Therefore, the estimators are random, but the parameters are considered fixed in the repeated samples, and they are not assigned distributions. On the other hand, in the Bayesian framework, probability is defined as the degree of belief and may depend on qualitative and/or quantitative as well as objective and/or subjective information. As a result the parameter is assigned a subjective probability distribution and is treated as if it is random.

In the case where s is unknown, we have to specify a prior for s, which leads to a joint prior for q. We assume uniformity on $\ln s$ rather than on s to ensure that scale parameters to lie between 0 and ∞. The most commonly used prior in this case is taking $\ln s$ to be uniformly distributed over the interval $-\infty < \ln s < \infty$. This is considered a non-informative prior because every possible value in the parameter space is an equally likely outcome. Therefore, the prior distribution for $p(\ln s) \propto constant$. When we transform $p(\ln s)$ to $p(s)$,

$$P(s) = P(\ln s)\left|\frac{\partial \ln s}{\partial s}\right| = constant\frac{1}{s} \propto \frac{1}{s} \tag{1.8}$$

The result (1.8) is the Jeffreys' prior for s.

$$P(s) \propto s^{-1} \tag{1.9}$$

Jeffreys' prior is not specific to s; it can be employed whenever a non-informative prior is needed. Jeffreys' prior for the scale parameter is proportional to the

square root of the determinant of the information matrix, treating the scale and location parameters as independent. A proper prior is defined as a distribution that does not depend on the data and integrates to one. Jeffreys' prior is improper since it does not integrate to one, but may yield posterior densities that do integrate to one and are therefore proper. When a non-informative prior is proper, the posterior function is proper and the Bayesian estimator is always proper.

Using non-informative priors for the parameters, we get the following joint prior by multiplying (1.6) and (1.9)

$$P(q) = P(s)P(b) \propto s^{-1} \tag{1.10}$$

1.2.2 Analysis with Informative Priors

Economists usually have ideas about the likely values of parameters. In this case, informative priors on the entire parameter vector q, or on a specific portion of q (while assigning a non-informative prior to the rest of the vector) are appropriate. For example, if the simple linear model is a representation of the consumption function, then y_t is consumption, x_{2t} is the income in period *t*, and b_2 is the marginal propensity to consume (MPC). The investigator knows from economic theory that 0 < MPC< 1. Incorporating this information as a prior can only improve the estimation process[4]. Non-informative priors may be assigned to the rest of the parameter vector if desired. The resulting joint prior will be

$$P(q) = P(b_1)P(b_2)P(s) \propto UNI(0,1)s^{-1} \tag{1.11}$$

[4] It is only appropriate to mention the following comment made by William H. DuMouchel to Jaynes (1985): "I agree strongly with Professor Jaynes that the real opportunities for Bayesians lie in the use of informative priors. How else could we hope to do better than frequentists? To use noninformative priors is, basically, to play on their turf."

where $UNI(0,1)$ is the prior distribution which assumes values of b_2 are equally likely between the interval 0 and 1, and non-informative priors are used for s and b_1.

1.3 Posterior Density Function

Based on Bayes' theorem, the data y affects the posterior through the likelihood function $f(y|q)$. The selection of the likelihood will depend on the data. The nature of the data generation process may indicate, for example, a normal, Poisson or logistic distribution.

The posterior distribution obtained by combining the prior and the likelihood functions may be algebraically convenient if both of the distributions were from the same family of density functions. When the prior and the posterior distributions belong to the same class of distributions, the normal family in this case, the prior is called a conjugate prior.

For example, if the investigator has a normal likelihood, the normal prior is said to be a natural conjugate prior, which leads to a posterior with the same functional form. On the other hand a prior with a different functional form may lead to a very complicated posterior function that was virtually impossible to deal with before the MCMC methods were available.

1.4 Inference via Markov Chain Monte Carlo Methods

Obtaining joint posterior functions is not difficult for simple models or conjugate families since all that is required is the product of all priors and the likelihood function. Even marginal posterior functions that are obtained by integrating out the nuisance parameter(s) can be found in lower dimensional cases. However, for more complicated problems and nonconjugate priors, making inferences from the posterior distribution

becomes very burdensome. Calculating or even approximating the marginal posterior densities requires high dimensional integrals or sampling from very complicated posterior functions. This is the main factor for the lack of empirical research using the Bayesian approach until last decade. There have been several different approaches to performing the required numerical integrations. Implementation of most of these methods requires numerical expertise as well as very sophisticated software.

The first of these approaches made use of the quadrature methods. Quadrature methods are helpful in integrating functions of the form

$$f(x)\times N(a, j)$$

where $f(x) = a_0 x^n + a_1 x^{n-1} + ... + a_{n-1} x + a_n$ *and* $N(a, j)$ is a normal density. However, it works when the dimension of the integral is no more than six. Some alternative approaches are the numerical integration using Cartesian product Gaussian quadrature rules (Davis and Rabinowitz 1984) and numerical integration via Monte Carlo methods such as importance sampling (Geweke 1988,1989) or sampling/importance sampling (Rubin 1987, 1988; Gelfand, Smith 1990). The objective of these non-iterative approaches is to obtain exact or approximate posterior distributions. These approaches fail in the higher dimensions and have proven to be less efficient than their iterative counterparts. In addition to numerical approaches, analytical approximations are also applied to calculate the marginals and the expectations. Laplace's method (DeBruijn 1961) and its extensions have received a lot of attention. These approximations are usually based on normal kernel expansions and often require two function maximizations.

The introduction of sample-based iterative methods that have revolutionized Bayesian econometrics have none of the shortcomings of the above-mentioned

approaches. They are able to handle high dimensional cases and are successful at exploring and summarizing posterior distributions, regardless of the family of the distribution. This new method of simulation that is broadly known as Markov Chain Monte Carlo (MCMC) is both feasible and provides sufficiently accurate results if used with care. In a surprisingly short period these MCMC methods, namely the Metropolis-Hastings algorithm (Metropolis, et. al 1953; Hastings 1970) and the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990) have emerged as extremely popular tools for the analysis of complex statistical models in the field of Bayesian econometrics.[5] Based on the problem at hand, the investigator can utilize either of these tools or a combination of them can be constructed. This makes it possible to sample from the complicated posterior distributions and/or compute posterior moments or any other inferential summary statistic. Calculation of the marginal posterior functions is an important part of Bayesian analysis, for the usual objective is to make inferences about individual parameters and/or provide graphs for those marginal posterior densities. MCMC methods facilitate such investigations.

1.4.1 Markov Chains: Definition and Concepts

A stochastic process is a collection of random variables. Let $x(t)$ be a random variable for each t in some set T. Since t often represents time, we refer the $x(t)$ as the state of process at time t. It might be the temperature at time t or the opening of the Dow Jones Industrial Average (DJIA) on day t. The state space of the stochastic process is the set of all possible values that the random variable $x(t)$ can take. A stochastic process may be discrete or continuous. We observe a discrete time stochastic process when t is a

[5] Shao (1999) provides a measure-theoretic formulation of Markov Chain Monte Carlo methods.

countable set of observations, $\{x(t);\ t = 0,1,2...\}$. If t is an interval of the real line, the stochastic process is said to be a continuous process, $\{x(t);\ t \geq 0\}$.

A stochastic process, $x(t)$, is said to be stationary if for all $t_1, \ldots, t_n$, the random vectors $x(t_1)\ldots x(t_n)$ have the same joint distribution as the random vector $x(t_1+s)\ldots x(t_n+s)$, where s can take any value. Ergodicity provides for the independence of the states separated by long intervals. That is, if we start observing the chain at time t as $t \to ¥$, continuation of this process will result in the same probability as $t \to ¥ + s$ for all s.

The conditions for stationarity may be hard to establish in practice. As a result, we define a weakly stationary process $x(t)$. If $E[x(t)] = c$ and $Cov[x(t), x(t+s)]$ does not depend on t, the process $x(t)$ is said to be weakly stationary. That is, if the first moment is fixed, and for all t the covariance between $x(t)$ and $x(s)$ depends only on $t-s$ and not on t, weak stationarity is satisfied.

Among many different stochastic processes, the Markov process is most suitable for our purposes due to its ease of calculation. The stochastic process, $\{x_t\}$, is a Markov chain if,

$$P\{x_{t+1} = j | x_0 = k_0, x_1 = k_1, \ldots, x_{t-1} = k_{t-1}, x_t = i\} = P\{x_{t+1} = j | x_t = i\} \text{ for } t = 0,1,\ldots$$

The stochastic process, $\{x_t\}$, has the Markovian property if the conditional probability of any future event (state) given the past and present event, is independent of the past event and depends only on the present event.

The probability that the process will make a transition from state i to state j is called a transition probability and is represented by P_{ij}, where

$$P_{ij} = P\{x_{t+1} = j | x_t = i\} \text{ for } t = 0,1,...$$

The basic assumptions of the Markov process are that the transition probabilities are non-negative and the process makes a transition to some state. The following conditions represent these two assumptions for all $i, j \geq 0$,

$$P_{ij} \geq 0 \text{ and } \sum_{j=0}^{\infty} P_{ij} = 1$$

The transition probabilities of a process are represented with a probability density function, which, is the specific distribution of the random variable, *x(t)*, conditional on the previous state, *x(t-1)*. This conditional probability density $p(x_t | x_{t-1})$ can be replaced by any function *p*(x)* where *p*(x)* $\propto$ $p(x_t | x_{t-1})$. The function *p*(x)* is the kernel of the transition density and is called the transition kernel.

N-step transition probabilities are defined as the probability that a process in state *i* will be in state *j* after *n* additional transitions. After defining one-step transition probabilities, P_{ij}, *n*-step transition probabilities, $P_{ij}^{(n)}$ are obtained using the Chapman-Kolmogorov equations (Ross 1993). These equations are,

$$\begin{aligned} P_{ij}^{(n+m)} &= P\{x_{n+m} = j | x_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{x_{m+n} = j, x_n = k | x_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{x_{n+m} = j | x_n = k, x_0 = i\} P\{x_n = k | x_0 = i\} \\ &= \sum_{k=0}^{\infty} P_{kj}^m P_{ik}^n \end{aligned}$$

Let P denote one-step and $P^{(n)}$ denote n-step transition matrices. Using the set of equations illustrated above, which are called Chapman-Kolmogorov equations, we can calculate the n-step transition matrix by simply taking the n*th* power of P such that $P^{(n)} = P^{(n-1+1)} = P \times P^{(n-1)} = P^{(n)}$. Thus, the n-step transition matrix can be obtained by multiplying the one-step transition matrix by itself n times.

Due to the stationarity of the Markov chains the n-step probabilities do not change over time. That is, $P\{x_{t+n} = j | x_t = i\} = P\{x_n = j | x_0 = i\}$ for t =0,1… This result follows from the constancy of one-step transition probabilities.

In some cases, some adjustments are needed to classify the process as a Markov chain. For instance, if we let the state at time t depend only on the conditions at time *t-2*, then the stochastic process would not be a Markov chain. However, it is possible to transform it to a Markov chain by stating that the state at time t is determined by the conditions at time *t-1* and *t-2*.

The Markov property and stationary transition probabilities are the basic requirements for a Markov Chain. Furthermore, a finite state Markov Chain requires a countable number of states and a set of initial probabilities; i.e., $P(x_0 = i)$ for all i.

In addition to these requirements, there are three properties that every Markov chain has to satisfy. These properties are irreducibility, aperiodicity, and ergodicity.

State j is accessible from state i if $P_{ij}^{(n)} > 0$ for some $n \geq 0$. $P_{ij}^{(n)}$ is the conditional probability of being in state j, n steps after starting at state i. If state j is accessible from state i, and state i is accessible from state j, the states are said to communicate. If all of the existing states of the chain communicate, then the state space is not partitioned. An unpartitioned Markov chain is called irreducible. The important implication of this

property is that it provides us with the reassurance that all states with positive probability can be reached from any starting point.

The second property of a Markov chain is aperiodicity. If the process enters a state only at time 0,2,4, … , this state *i* said to have period 2. On the other hand, if the process enters a state 0,1,2, … , this state is said to have period 1 and is called aperiodic. The chain is periodic with period d if all states are periodic with period $d > 1$ and aperiodic if all its states are aperiodic. This property ensures that the chain does not cycle through a finite number of states.

Irreducibility and aperiodicity are sufficient conditions for ergodicity (Tierney 1994) which forms the basis for MCMC methods. Given the ergodicity result is satisfied, the chain qualifies to be a Markov chain.

If the chain has a stationary distribution besides being ergodic, there are two very important results that follow. First, the *n*th iterate of the transition kernel, as $n \to \infty$, converges to the invariant distribution, *g(x)*. This invariant distribution is also called the equilibrium distribution, and it may be a posterior distribution or any other target distribution that we want to sample from indirectly. The invariance condition states that if x_t is distributed according to *g(x)*, then so will be all subsequent elements of the chain. If the drawings are made from $P^{(n)}\left(x_{t+1}|x_t\right)$, then for large *n*, the probability distribution of the drawings is the invariant distribution, regardless of the initial value (Chib and Greenberg 1996). Second ($P^{(2)}$), third ($P^{(3)}$), fourth ($P^{(4)}$), and eighth ($P^{(8)}$) step transition matrices can be calculated in simple cases to detect convergence. The convergence will be observed when the matrices of step *n-1* and *n* become identical. The

matrix at the point of convergence is the equilibrium (invariant) matrix. The rate of convergence will depend on the complexity of the problem at hand.

Second, averages of the functions evaluated at the sampled values converge to their expected values under the target density as $n \to \infty$. This result helps us calculate the moments of the functions of the parameters, even when the functions involve nonlinearities. In case of MCMC methods, the key is to create a Markov process whose invariant (equilibrium) distribution is the posterior distribution. That provides us with a sample of values from the posterior distribution, without sampling directly from the target distribution. Using the properties of Markov chain, it is possible to demonstrate that as the sample size increases, its distribution will converge to the distribution of the posterior density. The Gibbs sampler and the Metropolis-Hastings algorithms are two main methods to achieve this objective.

1.4.2 Metropolis and Metropolis-Hastings Algorithms

The Metropolis algorithm (Metropolis *et al*. 1953) was developed to investigate the equilibrium properties of large systems of particles in an atom. Hastings (1970) suggested a generalization of this algorithm and illustrated how to simulate normal and Poisson deviates. Metropolis-Hastings (M-H) is a general name of a family of algorithms that encompasses Metropolis and its generalization Metropolis-Hastings. Both algorithms create a Markov Chain with a specific equilibrium distribution, which is the target density. This allows us to sample from intractable posterior distributions where other known generators fail. The transition probability distribution should be constructed in such a way that the Markov chain converges to a unique stationary distribution that is the posterior distribution.

Metropolis and Metropolis Hastings (M-H) algorithms are used to sample from distributions that are very hard to sample from directly. In this algorithm, we draw values of the unknown parameters from an approximate distribution and then correct those draws to get closer to the target density, $g(x)$. The target density is the probability density function that we want to sample from. In the Bayesian framework the desired density takes the form of the posterior density, $g(x) = P(q \mid y)$. The methodology is very similar to that of importance sampling. Importance sampling is also an iterative algorithm that is used when we cannot sample directly from a target density, $g(x)$. One would sample from a density, $I(x)$, which is easy to sample from and approximates the target density, $g(x)$. Using the sampled value, an importance ratio is calculated based on the ratio of $g(x)/I(x)$. This ratio is then used to approximate the target density. The difference between importance sampling and the M-H algorithm is the fact that drawings in M-H algorithm are sequential, and the distribution of the draws depends on the previous value drawn. This is the property that makes the M-H method a Markov chain, whereas in importance sampling the distribution remains the same at each iteration. The improved M-H draws get closer to the target density at each stage.

In order to perform the algorithm we need a sequence of draws given a starting value x^0 that we draw from a starting distribution $\Pi(x)$. Drawing a value x from a candidate distribution (or jumping distribution) $Q(x, x^0)$ generates the next candidate value in the sequence. For the algorithm to be efficient, the jumping distribution should: (1) be easy to sample, (2) be easy to calculate the rejection rule from, (3) travel a

reasonable distance in the parameter space and (4) have a reasonable rejection rate. Once x has been generated, it is accepted with probability $a\,(x^0, x)$ where,

$$a\,(x^0, x) = \text{Min}\left[\frac{\Pi(x)\times Q(x, x^0)}{\Pi(x^0)\times Q(x^0, x)}, 1\right].$$

If the candidate is accepted, another value is sampled and evaluated based on the same criteria. This process continues until a rejection takes place. In case of the rejection of a candidate, the next sampled value is taken to be the current value and evaluated. However, even if the jump is not accepted, it still counts as an iteration in the algorithm. The resulting sequence converges in distribution to the equilibrium density that is the target density, which usually is the posterior density.

There is a small difference in the execution of the Metropolis and the Metropolis-Hastings algorithms; both algorithms create a sequence of random numbers whose distribution converges to our posterior distribution. The starting point of both algorithms is sampling starting values, x^0 from a starting distribution, $\Pi(x)$. However, for the Metropolis the jumping distribution has to be symmetric, that is $Q(x, x^0) = Q(x^0, x)$, unlike the Metropolis-Hastings where the symmetry assumption is relaxed. Once a symmetric jumping distribution has been selected, the next value in the sequence, x is then drawn from this jumping distribution. Since the distribution is symmetric, the ratio of densities boils down to

$$R = \frac{\Pi(x)}{\Pi(x^0)}$$

In this case the candidate is accepted with probability of min(R, 1). If it is accepted, then we set the $x^0 = x^t$. On the other hand, if the candidate is rejected, $x^0 = x^{t-1}$.

The Metropolis-Hastings algorithm is the generalization of Metropolis algorithm. The main advantage of M-H is its speed compared to that of the Metropolis algorithm. In M-H the symmetry requirement is relaxed and R changes to be,

$$R = \left[\frac{\Pi(x) \times Q(x, x^0)}{\Pi(x^0) \times Q(x^0, x)} \right]$$

Example: Let our target density be a bivariate normal distribution. Therefore $P(q|y) = N(q|0, I)$. The jumping distribution in this case may be a bivariate normal that $Q(x, x^0) = N(q^*|0, I)$.

The following is a pseudo-code for this example where Metropolis algorithm is used:

SAMPLE FROM A BIVNOR x^0; /*to be used as starting values*/

DO m = 1 to M; /* Start SIMULATION */

S = [ ]; /*storage matrix for gth iteration*/

SAMPLE FROM A BIVNOR *x;* /* the candidate value*/

CALCULATE R = BIVNOR (*x*) / BIVNOR (x^0);

STORE *x*;

U = UNI (0,1);

IF U ≤ R THEN SET x^0 = *x*; /*return to the top of the loop*/

The algorithm continues until the convergence has been reached. The resulting sample, upon convergence, is a sample from the target density.

1.4.3 The Gibbs Sampler

A special case of M-H family of algorithms is Gibbs sampler, in which the acceptance ratio is always one. That is, every jump is accepted in this algorithm. This

result surfaces because of the definition of the jumping distribution where the jumps takes place only along the single subvector in question and does that with its conditional density given. Gibbs sampling is a way of approximating the posterior distribution in cases where it is intractable to do so with analytical and numerical methods. In its simplest form the Gibbs sampler works by constructing an algorithm to sample from a multivariate distribution given only the full conditional distributions. Geman and Geman (1984) first developed the algorithm for simulating posterior distributions in image construction. The algorithms designed by Geman and Geman are very similar to those of Metropolis, et. al (1953) and Hastings (1970). Tanner and Wong (1987) and Gelfand and Smith (1990) extend the Gibbs sampler.[6]

The Gibbs sampler is especially helpful for such complex Bayesian models as hierarchical Bayesian models, data augmentation, Bayesian applications of censored models and models with missing data. The recent popularity of the Gibbs sampler enabled us to observe a lot of applications in these areas of research. Gelfand and Smith (1990) and Gelfand, Hills, et. al (1990) are examples of an application of Gibbs sampler for developing marginal posterior densities for Bayesian problems which were previously inaccessible. The earliest applications of the Gibbs sampler to censored regression are Wei and Tanner (1990), Chib (1992), and Geweke (1992). Linear regression models with constrained parameters have been researched prior to the utilization of the Gibbs sampler by Judge and Takamaya (1966), Lovell and Presott (1970), and Davis (1978). These applications include traditional as well as the Bayesian treatments. Geweke (1986, 1995c) used posterior simulators and the Gibbs sampler to facilitate constraint

[6] There are also some hybrid versions of the Gibbs sampler and the Metropolis-Hastings algorithm such as the work of Tierney (1994).

applications. The seemingly unrelated regression (SUR) model of Zellner (1962) has been applied extensively in econometrics. Blattberg and George (1991) and Percy (1992) have studied the Gibbs sampling algorithm for the model introduced by Zellner. Chib and Greenberg (1995) and Min and Zellner (1993) have considered variations of SUR model. Probit models and extensions to panel data were some of the applications by Geweke, Keane, and Runkle (1994b) and McCulloch and Rossi (1995b).

As far as simultaneous equation models are concerned, there have been a great deal of Bayesian applications using a variety of priors, such as the works of Richard (1973) and Rothenberg (1975), as well as some applications with improper priors. These applications include the works of Chao and Phillips (1994) and Kleibergen and van Dijk (1994). The Gibbs sampler provides a practical solution for some of these applications, but requires further work for the rest of the simultaneous equation models.

Although most of the applications have been in a Bayesian framework, the Gibbs sampler may also be applied to the classical approach as Tanner (1991) illustrated with many such examples. These examples include the application of Gibbs sampling for likelihood functions in cases of missing observations (Gelfand and Carlin 1993), and the use of the Gibbs sampler for location of the modes of the likelihood function and simulation of sufficient statistics within the conditional frequentist paradigm (Geyer and Thompson 1992).

The Gibbs sampler is a posterior simulator that provides a scheme to generate random variables without having to calculate the marginal density. The algorithm itself uses a Markovian updating scheme that starts with an arbitrary set of starting values and conditional distributions. The ergodicity condition requires that the supports of the

conditional densities not be separated into disjoint regions. The objective of any Markov Chain Monte Carlo simulation is to come up with a transition density whose invariant distribution is the target density. In most cases the target density is the posterior distribution. The Gibbs sampler defines the transition density as the product of the set of full conditional densities starting from some initial value. Then the scheme updates full conditional densities given the previous rounds' values of the conditioning parameters. As the length of the simulation goes to infinity, and under the condition of the ergodicity of the Markov chain, the Gibbs sampler results are reliable. However, convergence and model sensitivity have to be checked for healthy results. Once the sample is generated, the collection of all the simulated draws can be used to summarize the target density by graphics and quantiles as well as moments be computed. Expectations can be calculated by taking a simple average of the function over the simulated draws and expected values of the functions of the parameters can be calculated with small modifications to the algorithm. Posterior predictive simulations of the unobserved outcome can be obtained by simulating conditional on the drawn values of q .

The initial step of the Gibbs sampler scheme deals with the sampling of the starting values $q^0 = (b_1^0, b_2^0, s^0)$ from an arbitrary starting distribution. Then, random drawings are successively made from the full conditional densities. Draws of b_1^1 from the full conditional density $\left[b_1 | b_2^0, s^0\right]$, b_2^1 from $\left[b_2 | b_1^1, s^0\right]$, and s^1 from $\left[s | b_1^1, b_2^1\right]$ are made. This completes the first iteration of the scheme. The first iteration yields the first sample of parameters, $q^1 = (b_1^1, b_2^1, s^1)$. The second iteration yields the second round of the parameters, $q^2 = (b_1^2, b_2^2, s^2)$. b_1^2 is drawn from $\left[b_1 | b_2^1, s^1\right]$, b_2^2 from $\left[b_2 | b_1^2, s^1\right]$,

and s^2 from $\left[s \middle| b_1^2, b_2^2\right]$. After *t* rounds of iterations, the joint sample of $\left(b_1^t, b_2^t, s^t\right)$ will be reached from the following individual distributions,

$$b_1^t \sim p\left(b_1 \middle| b_2^{t-1}, s^{t-1}, data\right)$$

$$b_2^t \sim p\left(b_2 \middle| b_1^t, s^{t-1}, data\right)$$

$$s^t \sim p\left(s \middle| b_1^t, b_2^t, data\right)$$

Geman and Geman showed that $\left(b_1^t, b_2^t, s^t\right) \sim \left(b_1, b_2, s\right)$ as $t \to \infty$. Therefore at t^{th} iteration we are going to arrive at the joint density of the unknown parameters for large enough *t*. The Gibbs sampler may be applied to a wide range of problems. The following examples are for the linear regression model.

We can consider the linear regression model in $y = Xb + e$ where $e \sim N(0, s^2)$ with the following conjugate priors of $vs^2/s^2 \sim c^2_{(v)}$ and $b|s^2 \sim N(\underline{b}|H_b^{-1})$, where $H_b^{-1} = s^2\left(X'X\right)^{-1}$. $\left(s^2\right)^{-(v+2)/2} \exp\left(-vs^2/2s^2\right) \exp\left[-\left(\frac{\left(b-\underline{b}\right)' H_b \left(b-\underline{b}\right)}{2}\right)\right]$

The likelihood function for the normal data is the form $\left(s^2\right)^{-T/2} \exp\left[\frac{-\left(y-Xb\right)'\left(y-Xb\right)}{2s^2}\right]$ or equivalently

$\exp\left(-vs^2/2s^2\right)\exp\left[-\frac{\left(b-b\right)'X'X\left(b-b\right)}{2s^2}\right]$ where $b = \left(X'X\right)^{-1}X'y$, $v = T-k$, and

$s^2 = v^{-1}(y - xb)'(y - xb)$. The joint posterior kernel is the product of the prior kernel and the likelihood function. However, economists are usually interested in the marginal posterior functions rather than the joint. Therefore, instead of integrating out the nuisance parameters, we may utilize the Gibbs sampler and get the marginal posterior distribution for b_1, b_2, s^2. For this purpose, after drawing the initial values, we sample from the full conditional densities for $vs^2/s^2 \sim c^2_{(v)}$ and $b|s^2 \sim N(\underline{b}|H_b^{-1})$. After t rounds we will obtain (b^t, s^t) which will converge to the joint distribution of (b, s^2) as $t \to \infty$. This is an easier way to obtain the marginal densities and make inferences even for this simple conjugate case.

The following pseudo-code is added to help facilitate the Gibbs sampler in this example:

SAMPLE $b^0, (s^2)^0$ or LET $b^0, (s^2)^0$ = OLS estimates; /* Starting Values */

DO m = 1 to M; /* Start SIMULATION */

S = [ ] /* storage matrix for parameters */

$\underline{b} = \left(H_0 + \sum X_t' \Omega^{-1} X_t\right)^{-1} \left(H_0 b_0 + \sum X_t' \Omega^{-1} y^t\right)$ /* H_0 : prior precision matrix */

$C = choleski\left[H_0 + \sum \left(X_t' \Omega^{-1} X_t\right)^{-1}\right]$; /* calculate choleski decomposition */

$U = N(0, I_k)$; /* sample a uniform rv */

OUTPUT $b = \underline{b} + C'U$; /* calculate b */

CALCULATE $s^2 = v^{-1}(y - xb)'(y - xb)$; /* Given the most recent value of β */

$c^2 = U'U$; /* sample chi-squared with v df*/

$invc^2 = 1/c^2$;

$s^2 = vs^2 / invc^2$; /* calculate s^2 */

STORE b, s^2;

SET $b^0 = b$, $\left(s^2\right)^0 = s^2$; /*return to the top of the loop*/

Note: Albert and Chib (1996) and Geweke (1995) are exceptional sources for more examples and additional computer-pseudo codes.

There are some important implementation issues associated with the Gibbs sampler. First, the investigator has to make sure the iterations have proceeded long enough, otherwise those simulations may be drastically misrepresenting the target distribution. Therefore, checking for convergence is one of the key issues. The investigator has to be aware if and when the algorithm is converging. Different convergence criteria that have been employed for M-H can also be utilized for the Gibbs sampler.

Second, even when convergence has been reached, the early iterations still resemble the starting approximation rather than the target density. Even for long runs, how much the initial series is affected by the starting distribution is difficult to determine. To fix that problem, we suggest discarding the beginning phases of the sequence, which is called the burn-in period of the sampler, and keeping the rest for our purposes.

Third, after the detection of convergence in the simulation experiment, it is suggested to drop every *k*th draw in the iteration with the purpose of breaking the possible interdependence between the draws. This process provides approximately independent draws from the target distribution. However, the resulting inefficiency from discarding some of the simulated data is not desirable. For instance, k=2 implies the

deletion of 50% of the simulated data after convergence. Gelman, et. al (1995) postulate that the posterior intervals obtained from such simulation quantiles would not be reliable.

Fourth, some researchers recommend making the inference based on *p* different simulated sequences of *n* runs, rather than a single sequence. This method is called the multiple path method. The investigator may run *p* parallel simulations and keep the last value in the sequence. This way it is possible to have an independent sample from the posterior distribution. However, since only last value from each of the *p* runs is used, instead of *pn* data points, the investigator makes use of only *p*. Therefore, the main disadvantage of the multiple path method is its inefficient use of data. On the other hand, the multiple path method may be helpful in the convergence check. The parallel simulations constructed before and after convergence may help detect the convergence by comparing the variances within and between the sequences (Gelman and Rubin 1992). If the variance within each sequence is much smaller than the variance between the sequences, it is an indication that the convergence is yet to be obtained.

The fifth important issue to consider when implementing Gibbs sampler is blocking. In some cases, it may be better to work with a complete breakdown of θ into its components, whereas some other cases may require that the parameter space be divided into subsets. The important question of whether or how to block the parameter space is determined by the correlation among the individual parameters. If highly correlated scalar components are treated individually, the convergence slows down considerably. This slow convergence is a result of autocorrelations that decay only slowly. However, when those correlated scalars are blocked together, the convergence becomes faster. One disadvantage of blocking is that it requires drawings from a

multivariate distribution as opposed to drawing from a univariate distribution. For example, in the classical linear regression model, two blocks are used. One block consists of the unknown coefficients and the variance of the error term is blocked separately.

A sixth consideration is data augmentation which is a very powerful tool that has been introduced by Tanner and Wong (1987). Full conditional distributions can be constructed by introducing latent and/or missing variables. In the classical regression framework, missing data can be added to sampler. Other examples include the works of Albert and Chib (1993a), McCulloch and Rossi (1994), Geweke, Keane, and Runkle (1994), and Wei and Tanner (1990). Probit and Tobit models where the observed data are augmented by the latent data to obtain the observed posterior are examples of data augmentation. Also, the idea of data augmentation can be applied to likelihood estimation with missing data models as well. The data set may be augmented for the missing data with the use of the Gibbs sampler.

Lastly, the investigator may have a conditional distribution that is hard to sample from. In such a case the sampling may be done through M-H algorithm. In such a case, we need a M-H algorithm within each Gibbs sampler. This method has been used and popularized by Muller (1991).

Despite the recent overwhelming enthusiasm about the Gibbs sampler, there are still a lot of hybrid models and variations of these models that need to be addressed. The Gibbs sampler provides us with a powerful tool to evaluate models and their extensions in a new light. Its ease of implementation is one of the other important issues that gave rise to its current and increasing popularity. These include applications such as

hierarchical models (Gelfand, Hills, Racine-Poon, and Smith, 1990), dynamic models

(Carlin, Polson, and Stoffer, 1992) mixed models (Gamerman, 1997) and hybrid models

(Besag, and Green, 1993, Muller, 1991).

1.5 References

Albert, J., and S. Chib (1996) "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo" in *Advances in Econometrics*, (eds., T. Fomby and R.C. Hill), 11, 3-24. Jai Press.

Albert, J., and S. Chib (1993a) "Bayesian Analysis via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts" *Journal of Business and Economic Statistics*, 11, 1-15.

Albert, J., and S. Chib (1993) "Bayesian Analysis of Binary and Polychotomous Response Data" *Journal of American Statistical Association*, 88, 669-679.

Besag, J. and Green, P. J. (1993) "Spatial Statistics and Bayesian Computation" *Journal of Royal Statistical Society, Series B*, 55, 25-37.

Blattberg, George (1991) "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations" *Journal of the American Statistical Association*, 86, 304-315.

Berger J. (1986) "Bayesian Salesmanship," in ***Bayesian Inference and Decision Techniques***, (eds., P. Goel and A. Zellner), Elsevier Science Publishers**.**

Carlin, B. P., Polson, N. G., and D. S. Stoffer (1992) "A Monte Carlo Approach to Non-normal and Nonlinear State-Space Modeling" *Journal of the American Statistical Association*, 87 493-500.

Chao, J.C. and P.C.B. Phillips (1994) "Bayesian Posterior Distributions in Limited Information Analysis of the Simultaneous Equations Model" Yale University Cowles Foundation Working Paper.

Chen, M., and Q. Shao (1998) "Monte Carlo Methods for Bayesian Analysis of Constrained Parameter Problems" *Biometrika*, 85, 73-87.

Chib, S., (1992), "Bayes inference in the Tobit Regression Model" *Journal of Econometrics*, 51, 79-99.

Chib, S., and E. Greenberg (1996) "Markov Chain Monte Carlo Simulation Methods in Econometrics" *Economic Theory*, 12, 409-31.

Chib S., and E. Greenberg (1995) “Understanding the Metropolis-Hastings Algorithms” *American Statistician,* 49, 327-35.

Davis, W. W. (1978) “Baysian Analysis of the Linear Model Subject to Linear Inequality Constraints” *Journal of the American Statistical Association,* 73, 573-579.

Davis, P. J. and P. Rabinowitz (1984). *Methods of Numerical Integration.* New York:Academic Press.

DeBruijn N. G. (1961) *Asymptotic Methods in Analysis*. Amsterdam, Netherlands: North-Holland.

Dorfman, J. H. and C. S. McIntosh (1999) “Imposing Inequality Restrictions: Advantage Bayesians” presented at American Agricultural Economics Association Meetings.

Dorfman, J. H., (1997) *Bayesian Economics Through Numerical Methods: A Guide to Econometrics and Decision-Making with Prior Information.* New York: Spring-Verlag.

Freedman D. A. (1986) “Reply” *Journal of Business and Economic Statistics*, 4, 126-7.

de Finetti (1937) "Foresight: Its Logical Laws, Its Subjective Sources" *Studies in Subjective Probability*, (eds., G. Koch and F. Spizzichino). Wiley: New York.

Gamerman K. (1997) “Efficient Sampling from the posterior distribution in Generalized Linear Mixed Models” *Statistics and Computing,* 7, 57-68.

Gelfand, A.E., and B. Carlin (1993) “Parametric Likelihood Inference for Record Breaking Problems” *Biometrika* 80, 507-515.

Gelfand, A.E. and B. Carlin (1992) “Bayesian Inference for Hard Problems Using the Gibbs Bampler” in *Computing Science and Statistics Interface,* (eds., C. Page, and R. LePage), 29-37. New York:Springer-Verlag.

Gelfand, A.E., and A.F.M. Smith (1992) “Bayesian Statistics without Tears: A Sampling-Resampling Perspective” *American Statistician,* 46, 84-88.

Gelman, A., and D. B. Rubin (1992) “Inference From Iterative Simulation Using Multiple Sequences” *Statistical Science* 7, 457-72.

Gelfand, A.E., and A.F.M. Smith (1990) “Sampling based approaches to calculating marginal Densities” *Journal of American Statistical Association,* 85, 398-409.

Gelfand, Alan E., Hills, Susan E., Racine-Poon, A., and A. F. M. Smith (1990) “Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling” *Journal of the American Statistical Association*, 85, 972-85.

Gelman, A., Carlin, John B., Stern, Hal S., and D. B. Rubin (1995) *Bayesian Data Analysis.* Chapman & Hall.

Geman, S. and D. Geman (1984) "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12, 609-628.

Geweke, J. F. (1997) "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints" in *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface,* (ed.,E.M. Keramidas), 571-578.

Geweke, J. F., (1995a) "Monte Carlo Simulation and Numerical Integration" *Handbook of Computational Economics,* (eds., H. Amman, D.Kendrick and J. Rust), Amsterdam: North-Holland.

Geweke, J. F. (1995b) "Simulation-Based Bayesian Inference for Economic Time Series" inpreparation.

Geweke, J. F. (1995c) "Bayesian Inference for Linear Models Subject to Linear InequalityConstraints," in *Forecasting, Prediction and Modeling in Statistics and Econometrics: Bayesian and non-BayesianApproaches,* (eds., W.O. Johnson, J.C. Lee and A. Zellner). New York: Springer-Verlag.

Geweke, J. F. (1992) "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments" in *Bayesian Statistics 4,* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith). Oxford, UK:Clarendon Press.

Geweke, J. F. (1991) "Efficient Simulation from the Multivariate Narmal and Student-t Distributions Subject to Linear Constraints" *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571-78.

Geweke, J. F. (1989) "Bayesian Inference in Econometric Models Using Monte Carlo Integration" *Econometrica*, 57, 1317-1339.

Geweke, J. F. (1988) "Comment on Priorier: Operational Bayesian Methods in Econometrics" *Journal of Economic Perspectives*, 2, 159-66.

Geweke, J. F. (1986) "Exact Inference in the Inequality Constrained Normal Linear Regression Model" *Journal of Applied Econometrics,* 1, 127-141.

Geweke, J. F., and Keane, M. (1995) "An Empirical Analysis of the Male Income Dynamics" University of Minnesota Department of Economics Working Paper.

Geweke, J. F., and Keane, M. and D. Runkle (1994) "Alternative Computational

Approaches to Inference in the Multinomial Probit Model" *Review of Economics and Statistics*, 76, 4, 609-632.

Geweke, J. F., Keene, M., and D. Runkle (1994b) "Statistical Inference in the Multinormal Multiperiod Probit Model" Federal Reserve Bank of Minneapolis, Staff Report 177.

Geyer, C. J., and E. A. Thompson (1995) "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference" *Journal of the American Statistical Association*, 90, 909-20.

Geyer C. J. (1992) "Practical Markov Chain Monte Carlo" *Statistical Science*, 7, 473-511.

Griffiths, W.E. and D. Chotikapanich (1997) "Bayesian Methodology for Imposing Inequality Constraints on a Linear Expenditure Function with Demographic Factors", I, forthcoming.

Griffiths, W., Hill R. C., and C. J. O'Donnell (2001) "Including Prior Information in Probit Model". Unpublished paper.

Griffiths, W., Hill, R. C., and P. Pope (1987) "Small Sample Properties of Probit Model Estimators" *Journal of American Statistical Association*, 82, 929-937.

Hastings, W.K. (1970) "Monte Carlo Sampling Methods using Markov Chains and their Applications" *Biometrica* 57, 97-109.

Hill, R. C. (1987) Modeling Multicollinearity and Extrapolation in Monte Carlo Experiments on Regression in *Advances in Econometrics*, 6, 127-155. JAI Press.

Jaynes E. T. (1985) "Highly Informative Priors" in *Bayesian Statistics* 2, 329-52, (eds., J. M. Bernards *et al).* Amsterdam:North-Holland.

Jeffreys, A. (1967) *The Theory of Probability*. London:Oxford University Press (3nd edn.).

Jeffreys, A. (1961) *The Theory of Probability*. Cambridge:Cambridge University Press (2nd edn.).

Judge, G., W. Griffiths, R. C. Hill, H. Lütkepohl, and T. Lee (1985) *The Theory and Practice of Econometrics*.New York:Wiley. (2nd edn.)

Judge G. G. and T. Takamaya (1966) "Inequality Restrictions in Regression Analysis" *Journal of the American Statistical Association*, 61, 166-181.

Judge, G. G., and T.A Yancey (1978) "Inequality Restricted Estimation Under Squared

Error Loss," University of California Working Paper.

Kleibergen F.R. and H.K. van Dijk (1994) "On the shape of the likelihood/posterior of Cointegration Models" *Econometric Theory,* 10, 514-551.

Leamer, E. (1985) "Sensitivity Analysis Would Help" *American Economic Review,* 75, 308-313.

Lovell, M. C., and E. Prescott (1970) "Multiple Regression with Inequality Constraints: Pretesting Bias, Hypothesis Testing and Efficiency" *Journal of the American Statistical Association,* 65, 913-925.

McCulloch, R.E. and P.E. Rossi (1995) "An Exact Likelihood Analysis of the Multinomial Probit Model" *Journal of Econometrics,* 64, 207-240.

McCulloch, R.E. and P.E. Rossi (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model" *Journal of Econometrics,* 64, 207-40.

McMulloch, George, and P. E. Rossi(1991) "An Exact Likelihood Analysis of the Multinomial Probit Model" University of Chicago Working Paper 102.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and E. Teller (1953) "Equations of the state calculations by fast computing machines" *Journal of Chemical Physics,* 21, 1087-1092.

Min C. and A. Zellner (1993) "Bayesian Analysis, Model Selection and Prediction" in **Physics and Probability: Essays in Honor of Edwin T. Jaynes, (*eds., W.T. Grandy, Jr. and P.W. Milonni). New York: Cambridge University Press.***

Muller M. (1991) "Monte Carlo Integration in General Dynamic Models," Contemporary Mathematics, 115, 145-63.

Percy D. F. (1992) "Prediction for Seemingly Unrelated Regressions" *Journal of the Royal Statistical Society Series B,* 54, 243-52.

Richard J. F. (1973) *Posterior and Predictive Densities for Simultaneous Equation Models* Berlin: Spinger-Verlag.

Rosenkrantz R. D.(1981) *Foundations and Applications of Inductive Probability.* Atascadero: Ridgeview.

Rosenkrantz, S. (1992) "The Bayes Factor for Model Evaluation in a Hierarchical Poisson Model for Area Counts" University of Washington unpublished Ph. D. thesis.

Ross, S., (1993) *Introduction to Probability Models.* Academic Press (5^{th} edn.).

Rothenberg T. J. (1975) “The Bayesian Approach and Alternatives in Econometrics” in *Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, (eds., S. E. Fienberg and A. Zellner), 55-67. Amsterdam:North-Holland Publishing Co.

Rubin D. R. (1988) “Using the SIR Algorithm to Simulate Posterior Distributions” in *Bayesian Statistics 3,* (*eds.* J. M. Bernardo *et al*.). Oxford:Oxford University Press.

Rubin, D. R. (1987) “A Non-iterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Informa tion Are Modest” *Journal of the American Statistical Association,* 82, 543-6.

Savage L. J.(1954) *The Foundation of Statistics.* New York : J. Wiley & Sons,.

Tanner, M. A. (1991) “*Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions”*. New York:Springer-Verlag (2^{nd} edn).

Tanner, M. A. and Wong (1987) “The Calculation of Posterior Distributions by Data Augmentation” *Journal of the American Statistical Association,* 82, 528-549.

Tierney L. (1994) “Markov Chains for Exploring Posterior Distributions” *Annals of Statistics,* 22, 1701-62.

Wei, G. C. G., and M. A. Tanner (1990) “Posterior computations for Censored Regression Data” *Journal of the American Statistical Association,* 85, 829-839.

Zellner A. and D.S. Huang (1962) "Further Properties of Efficient Estimators for Seemingly Unrelated Regression Equations" *International Economic Review,* 2, 300-313.

Zellner, A. and C. K. Min (1995) “Gibbs Sampler Convergence Criteria” *Journal of the American Statistical Association,* 90, 921-7.

CHAPTER 2

THE RISK OF INEQUALITY-RESTRICTED PROBIT ESTIMATOR

2.1 Introduction

Traditional econometric models assume a continuous dependent variable. In the case of discrete or limited dependent variable models, the Ordinary Least Squares (OLS) estimator is still consistent, but the estimated probability that an event will occur given the set of regressors can lie outside the $(0,1)$ range. Discrete choice, also called qualitative response, models could take the form of a binary or multinomial choice structure, where the dependent variable might be mixtures of discrete and continuous outcomes. Early applications of qualitative response models dealt exclusively with cross-sectional data. In discrete-choice models of individuals, a maintained assumption is that each individual's random utility shock is an independent draw from the population distribution. More recently, there have been cases with purely time-series data in the macroeconomic literature that deal with models of multi-period expectations with discrete outcomes such as in Stock and Watson (1991) and Estrella and Mishkin (1997).

The binary choice model is one of the most widely used discrete choice models. This particular type of model is very relevant in economic analysis because economic units are usually in a position to make a decision where the outcome is dichotomous such as buying a car or not, buying a house or renting, joining a union or not, joining the labor force or not, and defaulting on a loan or not, among others. Some statistical models specify the probability of a discrete outcome as a function of independent variables and unknown parameters. A variety of models are possible depending on the choice of the

probability distribution for the residual term, which is termed as the link function in the econometrics literature. One of the most commonly used distribution functions is the standard normal distribution, which yields the probit model. Parameters of the probit model are estimated via the maximum likelihood estimator (MLE), which requires numerical maximization of the log likelihood function. On the other hand, the use of the logistic distribution for the error term results in the logit model. One should be cautious of the possibility of a misspecified link function, because this could lead to substantial bias in mean response estimates as illustrated in Czado and Santner (1992). Introducing a skewed distribution for the underlying latent variable, Chen, Dey and Shao (1999) develop a class of asymmetric link models for binary response data. An important contribution of this paper is that the skewed link model leads to a straightforward informative prior elicitation scheme based on historical data that yields proper priors.

Economic theory often translates into inequality restrictions on parameters in empirical studies. Some restrictions, such as monotonicity and concavity restrictions on the cost function, come from the economic theory, while other restrictions may stem from a particular model. In an empirical study, one key question is whether one should impose inequality restrictions on the parameters.

This chapter compares the performance of Bayesian and maximum likelihood point estimators in the context of the binary choice model with and without inequality restrictions imposed on the parameters. For these comparisons, we assume a quadratic loss function and use the posterior mean as the Bayesian point estimator. The comparisons are made using the mean square error criterion.

There have been numerous studies, such as Judge and Yancey (1978) that compare the performance of inequality-constrained regression with its unconstrained counterpart. These studies show that the introduction of inequality constraints

(1) introduces bias to the estimator,

(2) reduces the variance of the estimator, and

(3) reduces the risk of the estimator when the direction of the inequality constraint is correct.

An extended application of the binary choice model can be to incorporate inequality constraints into the estimation process of the probit model. Griffiths, Hill and O'Donnell (2001), in a Metropolis-Hastings algorithm setting, include inequality information on the signs of the coefficients to make Bayesian inference about probabilities and elasticities. The results are substantially different from those obtained using maximum likelihood estimation. Furthermore, they postulate that placing prior information on the choice probabilities, rather than the coefficients, can have a dramatic impact on the posterior probability density functions (pdfs) for the coefficients, the choice probabilities and the elasticities.

Previous work on the linear inequality constraint has been concentrated on the classical theory approach to inference. Such works include Judge and Takayama (1966), Zellner (1961), and Lovell and Prescott (1970). Only a very small number of works have been done on the Bayesian analysis of regression models with restricted parameter space. Lindley (1961), O'Hagan (1973), Davis (1978), Judge and Yancy (1978), Schmidt and Thomson (1982), Sedransk, Monahan and Chiu (1985), Greene and Seaks (1989),

Gelfand, Smith, and Lee (1992), Pace and Gilley (1993), Geweke (1995), Wan and Griffiths (1995), and Kleibergen (1997) are the most significant examples.

In economic models, sign or inequality parameter constraints are very common. In most cases the economic theory provides us with a priori information about the expected signs or ranges of the parameter coefficients. In these cases where we have prior information we may want to incorporate them in our statistical analysis. Bayesian inference, with Gibbs sampling, presents a simple and practical solution to incorporating those restrictions, compared to its classical inference counterpart. Geweke (1995) used the Gibbs sampler for linear models subject to inequality constraints. In this chapter, We incorporate inequality constraints into the Bayesian binary choice model. This method can be used in a binary choice content whenever we wish to employ sign or inequality constraints on the coefficients of the binary choice model.

Zellner and Rossi (1984) was one of the first papers in Bayesian literature to work on the binary quantal response models. Tanner and Wong (1987) introduced the data augmentation idea into the Bayesian and maximum likelihood literature. Wei and Tanner (1990) used data augmentation in censored regression model. Chib (1992) was the first paper to apply the Markov Chain Monte Carlo technique to a Tobit model. Albert and Chib (1993) developed exact Bayesian methods for modeling categorical response data using the data augmentation. Development of efficient simulation routines played a big role in the advancement of Bayesian methods for categorical data. Simulation tools made it possible to sample from the multivariate normal and student-t distributions subject to inequality constraints Geweke (1991). In a Bayesian framework, Geweke (1995) used the Gibbs sampler to impose linear inequality constraints on the coefficients of the normal

linear regression model. Griffiths and Chotikapanich (1997) imposed inequality restrictions on a linear expenditure system. More recently, Chen and Shao (1998) imposed parameter constraints in a hierarchical Bayesian model.

By combining the two strands of the literature we hope to illustrate the gains and losses associated with imposing inequality restrictions on both MLE and Bayesian estimators. To this end, we will review the probit maximum likelihood estimator in Section 2.2.1, the Bayesian methodology in Section 2.2.2, and the Gibbs sampler in Section 2.2.3. The designs and results of two Monte Carlo experiments will be explained in Section 2.3, and the conclusions are presented in Section 2.4.

2.2 Probit and Bayesian Estimators for Binary Choice Models

The probit estimator is widely used in applied work where the observable dependent variable is binary. The asymptotic properties of this estimator are well known. In addition, the finite sample properties have also been researched in Griffiths, Hill, and Pope (1987). Bayesian methods have become more popular after the development of computer technology. The binary choice model has been estimated using the Bayesian computational techniques.

In a univariate binary choice model, the observable dependent variable y_i will have dichotomous outcomes. The outcomes indicate whether or not some event occurred. The dependent variable will take on the value 1 if the event occurred and 0 otherwise.

The random utility model, in which economic units make choices to maximize their utilities, underlies the binary choice model. An event will occur with probability P_i only if the utility obtained from that event is more than the utility obtained from the other

event. The probit model can be derived based on an unobservable variable, called a latent variable in the econometrics literature.

Let y_i^* be the difference between the utilities obtained from the two alternatives. y_i^* is positive if the utility obtained from the first alternative is greater than that of the second alternative. Since we observe only whether or not the event occurred, we only observe the sign of y_i^*, which is a continuous random variable and assumes a linear function of some predetermined variables, x_i, with a stochastic error term.

$$y_i^* = x_i'b + e_i \tag{2.1}$$

where e_i are $IN(0, s^2)$.

One would observe only the sign of y_i^*, which determines the value of the observed binary variable, y_i,

$$y_i = \begin{cases} 1 & if\ y_i^* > 0 \\ 0 & if\ y_i^* \le 0 \end{cases} \tag{2.2}$$

In the probit model, the probability that $y_i = 1$ is

$$P_i = P(y_i = 1) = P(y_i^* > 0)$$

from the relation (2.1), we get

$$P_i = P[x_i'b + e > 0]$$

$$P_i = P[e_i > -x_i'b]$$

by symmetry,

$$P_i = P[e_i \le x_i'b]$$

$$P_i = P\left[\frac{e_i}{s} \le \frac{x_i'b}{s}\right]$$

assuming $s = 1$,

$$P_i = P\left[z_i \le x_i'b\right]$$

$$P_i = \Phi\left(x_i'b\right) \tag{2.3}$$

where x_i is a $(K \times 1)$ vector of nonstochastic predetermined variables, b is a $(K \times 1)$ vector of unknown parameters, and Φ is assumed to be the cumulative distribution function of a standard normal variable,

$$\Phi\left(x_i'b\right) = \int_{-\infty}^{x_i'b} \frac{1}{\sqrt{2p}} \exp\left[-\left(\frac{t^2}{2}\right)\right] dt \tag{2.4}$$

2.2.1. Maximum Likelihood Estimation of the Probit Model

The most widely used estimator for the probit model is the Maximum Likelihood Estimator (MLE). The MLE of the unknown parameters, b, is the vector of values that maximizes the likelihood function for a sample of n observations.

Since the probability P_i, of event E occurring is a function of some predetermined variables and unknown parameters, $P_i = \Phi\left(x_i'b\right)$, we can write the probability function for y_i in the form,

$$P\left(y_i\right) = \Phi\left(x_i'b\right)^{y_i}\left[1 - \Phi\left(x_i'b\right)\right]^{1-y_i}$$

Assuming a sample of n independent observations, the likelihood function is,

$$L = \prod_{i=}^{n}\left[\Phi\left(x_i`b\right)\right]^{y_i}\left[1 - \Phi\left(x_i`b\right)\right]^{1-y_i}$$

Since the likelihood function is the distribution of the data given the unknown parameters, we can also write the likelihood function as,

$$L = L(b|data)$$

Taking the log of the likelihood function will give us the log likelihood function that will be used in the maximization process,

$$\ln L = \sum_{i=1}^{n} y_i \ln[\Phi(x_i'b)] + (1-y_i)\ln[1-\Phi(x_i'b)]$$

The MLE is obtained by maximizing the log likelihood function with respect to b. The first order conditions are nonlinear; therefore the maximum likelihood estimates are obtained numerically. Throughout the paper, the Newton-Raphson algorithm, as used in the NLPNRA call of SAS/IML nonlinear optimization subroutine has been utilized. This algorithm uses this pure Newton step at each iteration when both the Hessian is positive definite and the Newton step successfully increases the value of the objective function.

In the case of the inequality constraints, the advantage takes advantage of the diogonal hessian matrices. Provided that the lower or upper bounds are specified, the subroutine uses the gradient and hessian and requires continous 1st and 2nd derivatives of the objective function to be inside the feasible region. In each iteration, a line search is done along the search direction to find an approxiamte optimum of the objective function using quadratic interpolation and cubic extrapolation. Optimum values that are outside of the feasible region are assigned the boundary value.

The algorithm is repeated until convergence is reached using the following iterative equation:

$$b_{t+1} = b_t - \left[\left(\frac{\partial^2 \ln L}{\partial b \partial b'}\right)\right]_{b_t}^{-1} \left[\frac{\partial \ln L}{\partial b}\right]_{b_t} \qquad (2.5)$$

The resulting MLE is consistent and asymptotically normal (Amemiya 1985) with an asymptotic distribution,

$$b_{MLE} \overset{asy}{\sim} N\left(b, \left[I(b)\right]^{-1}\right)$$

where

$$I(b) = E\left[\left(\frac{\partial \ln L(b)}{\partial b}\right)\left(\frac{\partial \ln L(b)}{\partial b'}\right)\right]$$
$$= -E\left[\frac{\partial^2 \ln L(b \mid y, x)}{\partial b \ \partial b'}\right]$$

MLE yields consistent and asymptotically efficient estimators. However, the estimator may not be reliable when the sample size is small (Albert and Chib 1993). The finite sample bias of MLE can be substantial (Griffiths, Hill and Pope 1987).

2.2.2 Bayesian Estimation

The first step in the Bayesian approach is to set up the joint prior density function $P(b)$. This joint density is the product of the marginal densities when we assume independence. The prior density reflects the nonsample or prior information. After the sample has been collected, Bayes' theorem is used to combine the sample and prior information. Bayes' theorem states that

$$P(b \mid y) = \frac{L(b \mid y)P(b)}{f(y)}$$

where $P(b \mid y)$ is the posterior density function, and $f(y)$ is the unconditional distribution of the sample. We can rewrite Bayes' theorem as

$$P(b \mid y) \propto L(b \mid y)P(b)$$

where $\propto$ denotes 'proportional to'. Given a normal likelihood and the marginal prior distribution, the generalized normal linear model has the following result:

$$P(b) \sim N\left(\tilde{b}, V_{\tilde{b}}\right)$$

where $\bar{b}$ and $V_{\bar{b}}$ are the prior mean and covariance matrix, respectively. Given the values of the latent variable, the latent variable model is exactly the normal linear model. Given the latent variable y^*, the posterior resulting from this prior and the likelihood is

$$P\left(b \mid y, y^*\right) \sim N\left(\tilde{b}, V_{\tilde{b}}\right)$$

where $\tilde{b} = \left(V_{\bar{b}}^{-1} + X'X\right)^{-1}\left(V_{\bar{b}}^{-1}\bar{b} + X'y^*\right)$ and $V_{\tilde{b}} = \left(V_{\bar{b}}^{-1} + X'X\right)^{-1}$.

In this study we use an uninformative prior for the unconstrained model. For all parameters, $P(b) \sim N(0, 100 \cdot I_k)$, where I_k is a $k \times k$ identity matrix. The variance for the prior is chosen to create a diffuse prior.

For the constrained Bayesian model, the inequality constraint is imposed via a truncated prior density. Bayesian inference is based on evaluating the posterior density function. The point estimates are the values that minimize the quadratic loss. Under a quadratic loss function, the Bayesian point estimates of b are the posterior means, given by

$$\int b \; P\left(b \mid y\right) db$$

Although the integral cannot be evaluated analytically in the probit model, we can compute the posterior means numerically.

2.2.3 The Gibbs Sampler

The Gibbs sampler is a method for computing the exact posterior density for the parameters. In its simplest form, the Gibbs sampler works by constructing an algorithm

to sample from a multivariate distribution given only the full conditional distributions. Geman and Geman (1984) first developed the algorithm for simulating posterior distributions in image construction. Their method and algorithms are very similar to those of Metropolis et. al (1953) and Hastings (1970). Tanner and Wong (1987), Gelfand and Smith (1990), Wei and Tanner (1990), Chib (1992), and Albert and Chib (1993) then extend the Gibbs sampler and its applications to different models.

In any model where the dependent variable is observable, the Gibbs sampler can be applied directly to the parameters of the model given the full conditional densities of the model. The Gibbs sampler will provide us with a sample from the exact posterior distribution. In other situations, such as the probit model, the data augmentation technique suggested by Tanner and Wong (1987) can be used to enlarge the parameter space with latent data. The idea of data augmentation is straightforward. In the probit model, if the underlying latent variables are known, one could iterate the Gibbs sampler directly. Therefore, the required conditional densities are the conditional density of the latent variable given the last sampled value of the unknown parameters, and the conditional density of the unknown parameters given the latent variable.

Bayesian methods, via the use of Markov Chain Monte Carlo methods utilizing data augmentation, provide us with the exact posterior distribution for the probit model. This allows us to calculate summary statistics on the unknown parameters and/or any function of these parameters. Data augmentation is very useful in the cases of missing values and latent variables. It refers to a process of augmenting the observed data with the missing or unknown observations. This will help us calculate the posterior distribution. In addition, Dueker (1999) shows how to use data augmentation methods to draw values of

the latent variable, whereupon its conditional heteroscedasticity can be addressed with regime switching techniques. To this end, he simplifies the estimation of the dynamic ordered probit model of Eichengreen et al. (1985) via the Gibbs sampler and its data augmentation.

To introduce this methodology we first need to define the prior distribution function, $P(b)$, that reflects our knowledge or ignorance on the unknown parameters, b. We will be able to introduce any constraint on the parameters at this stage. This will allow us to develop the posterior distribution function using the Bayes' theorem. Based on Bayes' theorem, the posterior function of the unknown parameters is proportional to the product of the prior distribution function and the likelihood function.

Given the likelihood function, $L = L(y|b)$ and the prior density function, $P(b)$, the posterior function based on the Bayes' theorem will be,

$$P(b\,|data) \propto P(b)\,L(data|b)$$

where the constant of proportionality is given by

$$f(data)$$

Let y and z be the observed and latent data, respectively. The next step in the process is to generate the latent sample. For this step we need to sample from a predictive distribution $P(z|y,b)$. As a result, the Gibbs sampler with data augmentation allows us to do the implicit integration and provides us with the exact distribution of the unknown parameters for this complex model.

If y and z were known, it would be easy to calculate the posterior, $P(b\,|y,z)$. Then by integrating out the latent data we would get the desired posterior density

function, $P(b|y)$. However, since latent data are not known, we will use data augmentation that allows us to generate multiple values of latent variable, z, by sampling from the predictive distribution, $P(z|y,b)$. The method consists of two-step iteration. In the first step, we sample from the predictive distribution, $P(z|y,b)$ given the initial values of the unknown parameters as suggested by Albert and Chib (1993) in conjunction with the data augmentation algorithm (Tanner and Wong 1987). These initial values can be the OLS or MLE estimates or values from an arbitrary distribution. The second step involves updating the posterior density of the unknown parameters with the sample from the first step, which is to sample from $P(b|y,z)$.

In the case of the Bayesian estimation of the probit model, data augmentation is used to augment the observed dependent variable with the unobserved value of y_i^* which depends on the underlying utility function. If the value of y_i^* were know, our model would mimic the linear regression model. Therefore, given the value of y_i^*, the properties of the coefficients are know, therefore are easy to sample from. Using the sampled value y_i^*, we can then construct the dependent variable y_i as,

$$y_i = \begin{cases} 1 & if\ y_i^* > 0 \\ 0 & if\ y_i^* \leq 0 \end{cases}$$

One needs to sample from two separate distributions, one distribution for sampling y_i^* that corresponds to $y_i = 1$, and another one that corresponds to $y_i = 0$.

Since the underlying distribution of the probit model is the normal distribution, values of the latent data are simulated from a truncated normal distribution. We will use

a right truncated normal distribution to sample the latent variable corresponding to $y_i = 0$ and a left truncated normal distribution corresponding to $y_i = 1$.

Given the simulated values of the latent data, the joint posterior distribution of the parameters can be computed. Draws from that posterior are used to simulate new latent data. The Gibbs sampler is used for this iterative process, which continues until convergence.

In short, the steps of the Gibbs sampler could be summarized as follows:

(1) Sample a latent variable based on the value of the binary variable given the initial values for the unknown parameters.

a) If the truncation is above 0 (i.e., if $y_i = 1$), then

$$f\left(y_i^* \middle| \beta, y\right) = \frac{f\left(y^*\right)}{\int_0^\infty f(t)\,dt}$$

b) If the truncation is below 0 (i.e., if $y_i = 0$), then

$$f\left(y_i^* \middle| \beta, y\right) = 1 - \frac{f\left(y^*\right)}{\int_0^\infty f(t)\,dt}$$

Draw values for the unknown parameters given the recently sampled value of the latent variable. This is equivalent to an OLS regression. Since given y^*, we can estimate $y^* = x_i'\beta + e$ very easily.

$\left(\beta \middle| y, y*\right) \sim N\left(\tilde{\beta}, (X'X)^{-1}\right)$ where $\tilde{\beta} = (x'x)^{-1}(x'y^*)$

Return to the first step and sample another latent variable given the most recent sampled values of the unknown parameters. After *t* iterations, the joint

sample of $\left(b_1^{(t)},\ldots,b_k^{(t)},y^{*(t)}\right)$ will be obtained. The first t_i iterations will be dropped from the sample to make the scheme robust to the initial value specification. Geman and Geman (1984) show that as t ® ¥, the joint density of $b^{(t)}$ approaches the joint distribution of b. Also, the marginal posterior densities for each parameter are obtained since $P\left(b_i^{(t)}\right) \rightarrow P\left(b_i \left| y, y^*\right.\right)$ as t ® ¥.

In some cases the prior distribution is considered to be a flat distribution, which reflects ignorance on the parameters. In this case, combining the likelihood function and the prior distribution results in a posterior distribution that is proportional to the likelihood function since the flat prior becomes a part of the constant term. Consequently, the resulting posterior will be identical to the log likelihood function. On the other hand, a prior density may be informative. When the economic theory provides us with some knowledge on the parameters, we could incorporate this information in the posterior calculations.

2.3 Monte Carlo Experiments

The objective of the experiment is to compare the performance of the Bayesian probit estimator with that of the maximum likelihood probit estimator under inequality constraints via the mean squared error (MSE), variance and bias measures. We examine the effects of varying degrees of specification errors on the performance of constrained maximum likelihood and Bayesian estimators for the probit model using Monte Carlo experiments. Estimator performance is evaluated by examining the finite-sample sampling-theory properties such as mean squared error (MSE), bias and variance.

Although these measures are irrelevant in Bayesian context, they are important to bring the two techniques on the common grounds to be able to compare them.

2.3.1 Design of the Experiment

We choose a sample size of *n*=100 and the true coefficients values of the vector b where $b = (-0.35, 0.5, 0)$. Two independent normal regressors of size *n* are used to estimate the unconstrained and constrained models. Nine different inequality constraints on the parameter space of b_3 are imposed to vary the specification error. 500 latent variables, y_i^* are generated by adding a $N(0,1)$ random error term, e_i, to the vector $x_i b$ for each of the models. The value of the binary variable y_i is obtained based on the sign of the latent variable.

The maximum likelihood point estimates are the values that maximize the likelihood function and are obtained by the Newton-Raphson optimization algorithm with OLS starting values in SAS/IML. The algorithm converged normally for all models and datasets. Robustness to different starting values is confirmed using the unconstrained probit starting values. The means of the sampling distributions are obtained by averaging over 500 observations for each model.

The point estimator for the Bayesian model is the posterior mean, which minimizes the expected quadratic loss function. A non-informative prior of $b \sim N(0,100)$ is used. This allows for a variance large enough to diffuse it, but not be unrealistically large. The Gibbs sampler generated 1100 samples from the posterior distribution, 100 of which are thrown away as part of the burn-in period. The posterior means for each dataset are calculated by averaging over the remaining 1000 samples in each dataset. Then, by averaging over 500 datasets, we obtain the means of the sampling

distribution for each model. The FORTRAN programming language is used for the Bayesian estimation. No formal convergence check is done, but the absolute changes in the parameter values are observed for possible nonconvergence. The same experiment has been repeated with a sample size n=50 to observe the impact of sample size on the performance of the parameter estimators.

MSE, bias and empirical variance are calculated for each Data Generating Process (DGP), where

$$MSE = \frac{\sum_{i=1}^{500} \left(\hat{\underset{\sim}{b}}_i - b \right)' \left(\hat{\underset{\sim}{b}}_i - b \right)}{500} \tag{2.7}$$

$$Bias = \bar{\hat{b}} - b \tag{2.8}$$

$$Cov\left(\hat{b} \right) = \frac{\sum_{i=1}^{500} \left(\hat{\underset{\sim}{b}}_i - \bar{\hat{b}} \right) \left(\hat{\underset{\sim}{b}}_i - \bar{\hat{b}} \right)'}{500} \tag{2.9}$$

$$Cov(MSE) = \frac{\sum_{i=1}^{500} \left(MSE_i - \overline{MSE} \right) \left(MSE_i - \overline{MSE} \right)'}{500} \tag{2.10}$$

2.3.2 Results

Table 2.1 reports the results for the unconstrained model for both of the estimation techniques. Column 1 contains the true values used to generate the simulated data. Column 2 contains the mean parameter estimates with their respective standard errors. Columns 3 and 4 contain the bias and the MSE of the parameter estimates. Figure 2.1 depicts the sampling distribution of the unconstrained ML and Bayesian

estimates for b_3. As can be seen from the plots, both distributions are similar and approximately normal with a slight skew to the left.

Table 2.1: Unconstrained Model (Sample size =100)

	True Value	Means	Bias	MSE
MLE(b_1)	-0.35	-0.3739 (0.1721)	-0.0239	0.0301 (0.0023)
MLE(b_2)	0.50	0.5458 (0.1286)	0.0458	0.0186 (0.0019)
MLE(b_3)	0.00	-0.0060 (0.0727)	-0.0060	0.0053 (0.0004)
BAY(b_1)	-0.35	-0.3792 (0.1764)	-0.0291	0.0319 (0.0025)
BAY(b_2)	0.50	0.5574 (0.1343)	0.0574	0.0213 (0.0022)
BAY(b_3)	0.00	-0.0064 (0.0744)	-0.0064	0.0056 (0.0004)

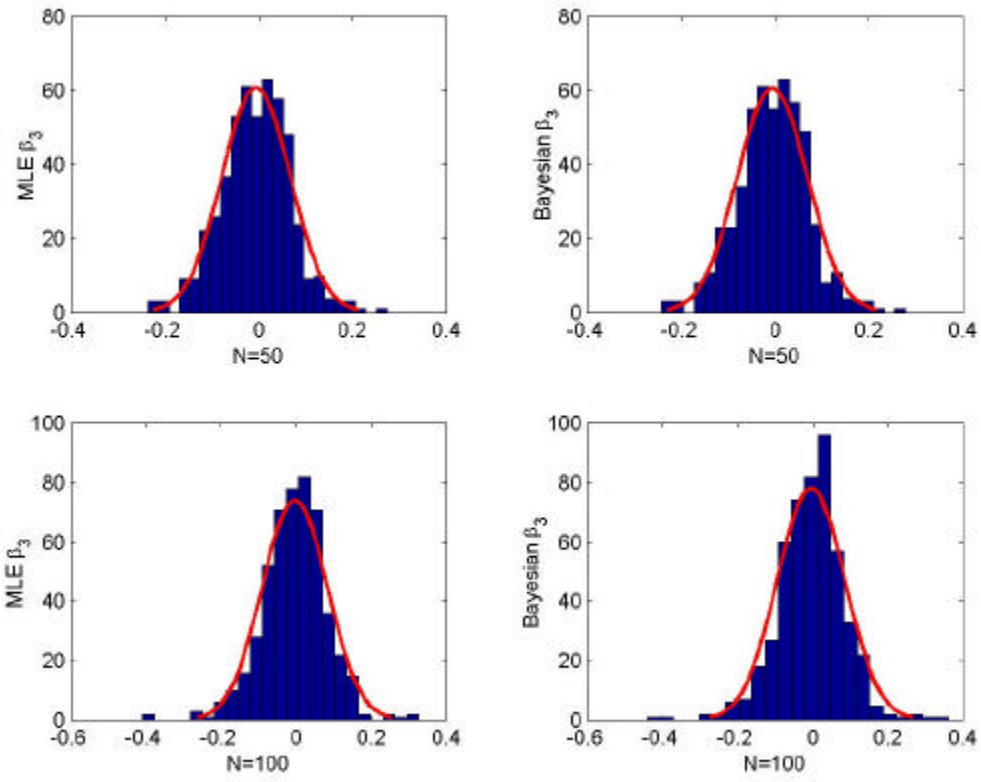


Figure 2.1: Sampling distributions

Not surprisingly, the unconstrained frequentist and Bayesian point estimates are quite similar. The small difference can be explained by the fact that with a non-informative prior on b_3, the MLE is essentially the posterior mode. The Bayesian estimator under the quadratic loss function is the posterior mean. Since the mean lies to the left of the mode when skewness is to the left, the

Bayesian method yields smaller estimates for b_3. Table 2.2 depicts the same information for a smaller sample of size 50.

Table 2.2: Unconstrained Model (Sample size =50)

	True Value	Mean	Bias	MSE
MLE(b_1)	-0.35	-0.3956 (0.2168)	-0.0456	0.0490 (0.0039)
MLE(b_2)	0.50	0.3249 (0.3011)	-0.1752	0.1212 (0.0048)
MLE(b_3)	0.00	-0.0042 (0.0863)	-0.0042	0.0074 (0.0007)
BAY(b_1)	-0.35	-0.4053 (0.2267)	-0.0553	0.0544 (0.0049)
BAY(b_2)	0.50	0.3407 (0.3204)	-0.1593	0.1278 (0.0052)
BAY(b_3)	0.00	-0.0045 (0.0894)	-0.0045	0.0080 (0.0008)

The results are very similar, except for the slightly better bias performance for the Bayesian estimator of b_2. However, in terms of overall risk, the MLE outperforms the Bayesian estimator even when the sample size is smaller.

We imposed nine different constraints on the parameter estimates b_3. These constraints allowed us to vary the specification error between –0.100 to 0.100. As the specification error gets more positive, the constraints get more inaccurate given the true parameter value of 0. For these constrained models, we observe upward biases in absolute value in both the MLE and Bayesian cases. For b_2 and b_3 the overall results

appear to be consistent with the findings of Zellner (1961). Tables 2.3, 2.4, and 2.5 show the bias, MSE and variance for the b_3 parameter estimates for different degrees of specification error, respectively.

Table 2.3: Bias of Bayesian vs. MLE Estimators

	Constraint is true					Constraint is false			
$b_3 \geq$	-0.100	-0.075	-0.050	-0.025	0.000	0.025	0.050	0.075	0.100
MLE(b_1)	-0.0225	-0.0216	-0.0203	-0.0189	-0.0169	-0.0153	-0.0143	-0.0140	-0.0147
MLE(b_2)	0.04417	0.0431	0.0418	0.0403	0.0390	0.0381	0.0383	0.0396	0.0424
MLE(b_3)	-0.0023	0.0009	0.0061	0.0140	0.0254	0.0401	0.0583	0.0796	0.1027
BAY(b_1)	-0.0275	-0.0284	-0.0281	-0.0271	-0.0266	-0.0261	-0.0252	-0.0236	-0.0230
BAY(b_2)	0.0569	0.0584	0.0582	0.0584	0.0583	0.0589	0.0593	0.0591	0.0590
BAY(b_3)	0.0060	0.0125	0.0212	0.0318	0.0450	0.0605	0.0782	0.0974	0.1182

Table 2.4: MSE of Bayesian vs. MLE Estimators

	Constraint is true					Constraint is falshe			
$b_3 \geq$	-0.100	-0.075	-0.050	-0.025	0.000	0.025	0.050	0.075	0.100
MLE(b_1)	0.0297 (0.0022)	0.0295 (0.0022)	0.0293 (0.0022)	0.0291 (0.0022)	0.0289 (0.0022)	0.0288 (0.0022)	0.0288 (0.0022)	0.0289 (0.0022)	0.0293 (0.0022)
MLE(b_2)	0.0181 (0.0019)	0.0179 (0.0019)	0.0175 (0.0018)	0.0172 (0.0018)	0.0168 (0.0017)	0.0166 (0.0017)	0.0164 (0.0017)	0.0165 (0.0017)	0.0167 (0.0017)
MLE(b_3)	0.0043 (0.0003)	0.0037 (0.0003)	0.0031 (0.0003)	0.0025 (0.0003)	0.0023 (0.0003)	0.0026 (0.0003)	0.0041 (0.0003)	0.0067 (0.0002)	0.01074 (0.0002)
BAY(b_1)	0.0312 (0.0023)	0.0316 (0.0024)	0.0317 (0.0025)	0.0316 (0.0025)	0.0314 (0.0024)	0.0315 (0.0024)	0.0312 (0.0023)	0.0312 (0.0024)	0.0313 (0.0024)
BAY(b_2)	0.0204 (0.0020)	0.0216 (0.0022)	0.0214 (0.0022)	0.0217 (0.0023)	0.0214 (0.0024)	0.0216 (0.0022)	0.0214 (0.0022)	0.0214 (0.0023)	0.0217 (0.0025)
BAY(b_3)	0.0034 (0.0003)	0.0030 (0.0003)	0.0027 (0.0003)	0.0027 (0.0003)	0.0032 (0.0003)	0.0045 (0.0003)	0.0067 (0.0003)	0.0099 (0.0002)	0.0142 (0.0002)

Table 2.5: Variance of Bayesian vs. MLE Estimators

	Constraint is true					Constraint is false			
$b_3 \geq$	-0.100	-0.075	-0.050	-0.025	0.000	0.025	0.050	0.075	0.100
MLE(b_1)	0.0293	0.0291	0.0290	0.0288	0.0287	0.0286	0.0286	0.0288	0.0291
MLE(b_2)	0.0162	0.0160	0.0158	0.0156	0.0154	0.0152	0.0150	0.0149	0.0149
MLE(b_3)	0.0043	0.0038	0.0031	0.0023	0.0016	0.0010	0.0006	0.0004	0.0002
BAY(b_1)	0.0306	0.0309	0.0310	0.0309	0.0308	0.0309	0.0307	0.0307	0.0309
BAY(b_2)	0.0172	0.0183	0.0181	0.0183	0.0181	0.0181	0.0179	0.0179	0.0183
BAY(b_3)	0.0033	0.0028	0.0022	0.0017	0.0012	0.0009	0.0006	0.0004	0.0003

The same information is captured in Tables 2.6, 2.7, and 2.8 for n=50.

Table 2.6: Bias of Bayesian vs. MLE Estimators

	Constraint is true					Constraint is false			
$b_3 \geq$	-0.100	-0.075	-0.050	-0.025	0.000	0.025	0.050	0.075	0.100
MLE(b_1)	-0.0420	-0.0406	-0.0387	-0.0362	-0.0334	-0.0303	-0.0273	-0.0246	-0.0226
MLE(b_2)	-0.1784	-0.1792	-0.1799	-0.1805	-0.1808	-0.1805	-0.1795	-0.1776	-0.1748
MLE(b_3)	0.0023	0.0058	0.0113	0.0194	0.0304	0.0444	0.0618	0.0821	0.1042
BAY(b_1)	-0.0532	-0.0504	-0.0488	-0.0465	-0.0439	-0.0419	-0.0386	-0.0357	-0.0328
BAY(b_2)	-0.1574	-0.1590	-0.1576	-0.1571	-0.1585	-0.1571	-0.1575	-0.1556	-0.1551
BAY(b_3)	0.0182	0.0262	0.0356	0.0474	0.0608	0.0761	0.0929	0.1112	0.1309

Table 2.7: MSE of Bayesian vs. MLE Estimators

	Constraint is true					Constraint is false			
$b_3 \geq$	-0.100	-0.075	-0.050	-0.025	0.000	0.025	0.050	0.075	0.100
MLE(b_1)	0.0472	0.0469	0.0466	0.0462	0.0459	0.0456	0.0456	0.0458	0.0462
MLE(b_2)	0.1190	0.1187	0.1184	0.1181	0.1179	0.1178	0.1177	0.1177	0.1179
MLE(b_3)	0.0053	0.0047	0.0040	0.0034	0.0032	0.0035	0.0049	0.0074	0.0113
BAY(b_1)	0.0543	0.0537	0.0526	0.0518	0.0520	0.0518	0.0516	0.0505	0.0508
BAY(b_2)	0.1289	0.1272	0.1290	0.1273	0.1268	0.1269	0.1257	0.1265	0.1258
BAY(b_3)	0.0042	0.0039	0.0039	0.0044	0.0054	0.0071	0.0096	0.0131	0.0177

Table 2.8: Variance of Bayesian vs. MLE Estimators

	Constraint is true					Constraint is false			
$b_3 \geq$	-0.100	-0.075	-0.050	-0.025	0.000	0.025	0.050	0.075	0.100
MLE(b_1)	0.0456	0.0453	0.0452	0.0450	0.0448	0.0448	0.0449	0.0453	0.0458
MLE(b_2)	0.0874	0.0867	0.0862	0.0857	0.0854	0.0854	0.0857	0.0864	0.0875
MLE(b_3)	0.0053	0.0047	0.0039	0.0031	0.0023	0.0016	0.0010	0.0007	0.0004
BAY(b_1)	0.0516	0.0513	0.0503	0.0497	0.0501	0.0501	0.0502	0.0494	0.0498
BAY(b_2)	0.1044	0.1022	0.1044	0.1028	0.1019	0.1024	0.1011	0.1025	0.1019
BAY(b_3)	0.0039	0.0033	0.0027	0.0022	0.0017	0.0013	0.0010	0.0007	0.0005

Since the inequality restrictions are imposed on b_3, there are significant changes in the bias, MSE and variance of that specific parameter estimate, but there are no

significant changes on b_1 or b_2 as the specification error changes on the estimation of b_3. These results have been depicted in Figures 2.2 and 2.3 for bias, in 2.4 and 2.5 for variance and 2.6 and 2.7 for MSE for sample sizes 100 and 50, respectively. The bias of b_3 for the unconstrained frequentist model is -0.0060, and it rises all the way up to 0.1027 when the most incorrect constraint is imposed (Figure 2.2).

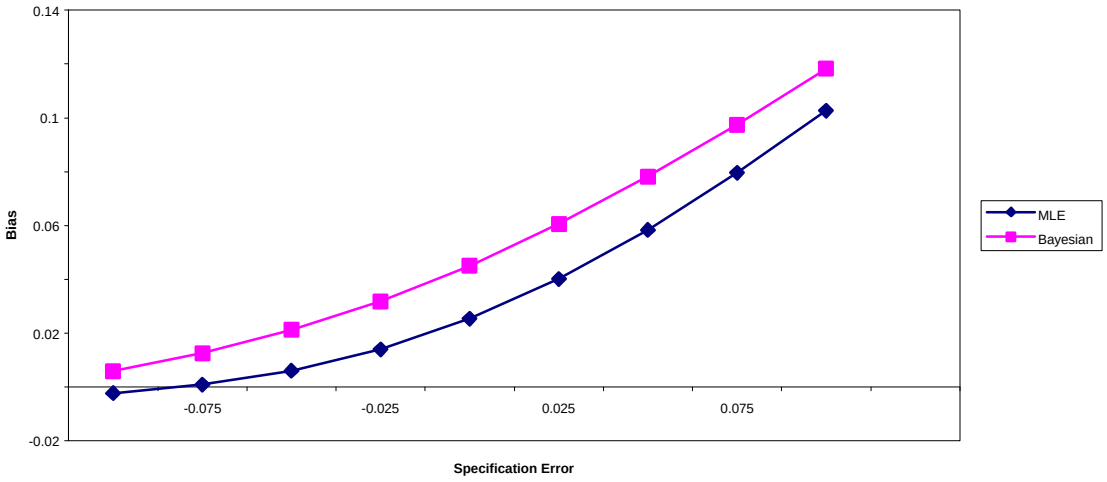


Figure 2.2: Bias Comparison for Beta3 (N=100)

When the sample size is 50, the bias of b_3 is less at –0.0042, but goes up to 0.1042 as the constraint gets more binding (Figure 2.3).

However, the variance of the estimator falls monotonically in the constrained model regardless of the magnitude or the correctness of the constraint. This result holds for both sample sizes and can be observed in Figures 2.4 and 2.5 for n=100 and n=50, respectively.

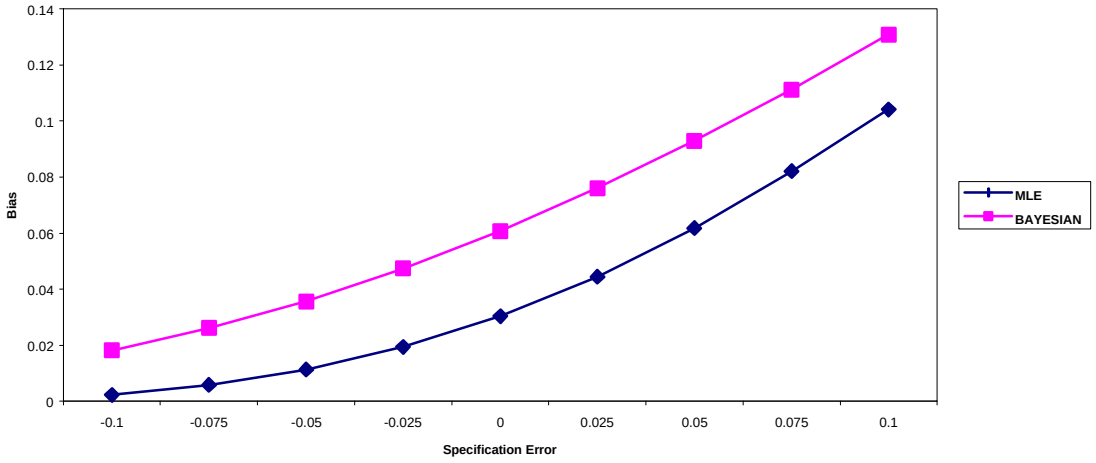


Figure 2.3: Bias Comparison for Beta3 (N=50)

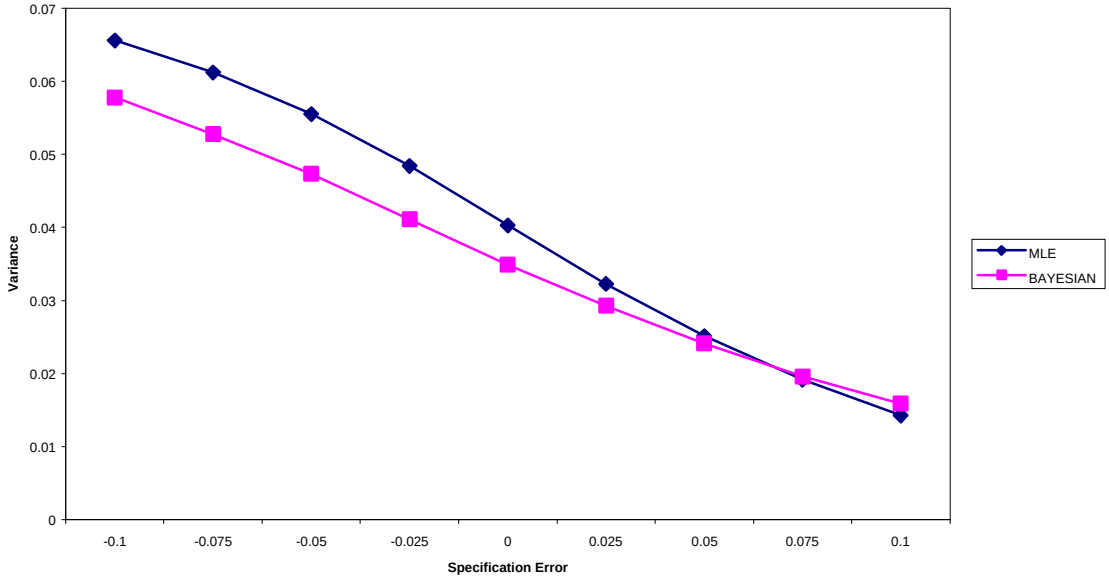


Figure 2.4: Variance Comparison for Beta3 (N=100)

This result also holds for the Bayesian model, with a larger upward bias as the constrained estimate rises to 0.1182 from -0.0064 for n=100 and from –0.0045 to 0.1304 for n=50. Although the bias of unconstrained b_3 is lower for both MLE and Bayesian methods for the smaller sample size, the faster increase in bias as the constraints get

binding indicates a less robust estimation. On the other hand, the variance gain for the Bayesian estimator is larger than it is for the MLE.

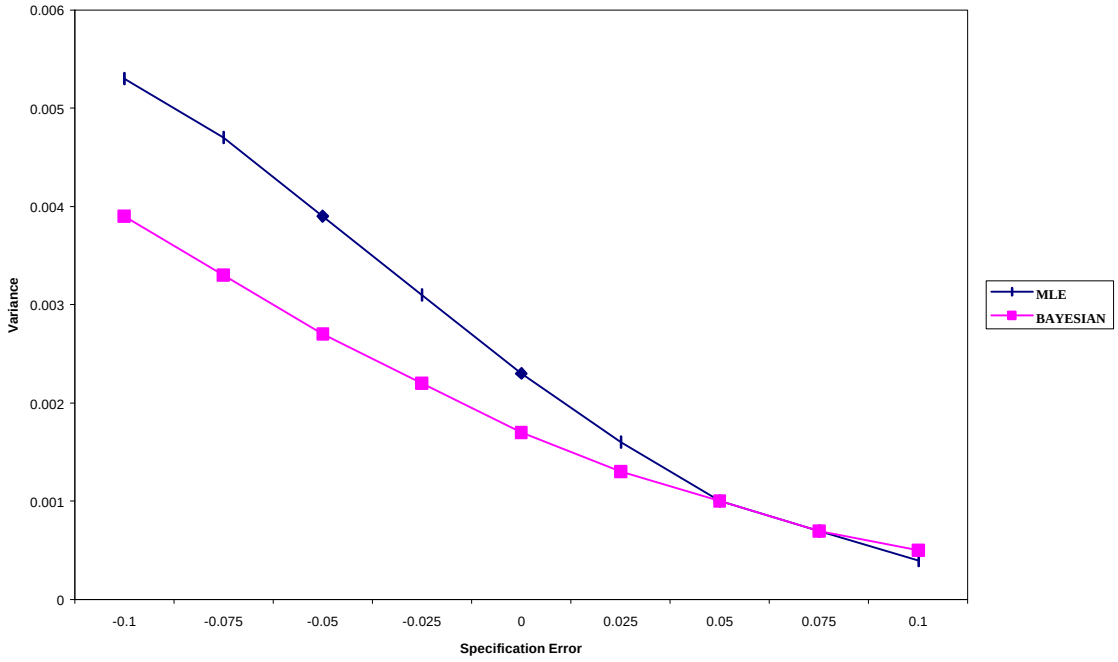


Figure 2.5: Variance Comparison for Beta3 (N=50)

In both cases, as expected, when the most binding constraints are imposed, the bias introduced in the estimation is the largest. Figures 2.2 and 2.4 compare the changes in the bias and variance for the two models as we vary the specification error. The trend in these figures is almost identical to those in Figure 2.3 and 2.5, which represent the bias and the variance for the estimation with smaller sample size. As can be observed from the plots, the bias of the Bayesian model is higher for every level of the specification error than the bias of MLE.

To understand the larger bias introduced by the Bayesian method, consider how the approaches would impose a non-negativity inequality restriction in a simpler

univariate model. Using the frequentist approach, negative estimates will be truncated to zero, leading to roughly a censored normal distribution for the estimator with a point mass at the constraint. Using the Bayesian approach, we slice away the portion of the posterior associated with negative values and scale the remaining density to integrate to one. Thus, the Bayesian approach leads to roughly a truncated normal distribution, which leads to larger coefficient values.

Hstograms for b_3 for varying degrees of specification error for both sample sizes are in Figures A.1-A.9. As we observe, the sampling distributions behave very similarly even when the sample size doubles. In all of the graphs, the first panel represents the sampling density obtained by the maximum likelihood estimation of b_3, and the second panel depicts the Bayesian density for a given inequality restriction. The graphs reveal patterns very similar to the results predicted above. The constrained frequentist model appears similar to a censored normal with mass associated with the boundary imposed on b_3, resulting in a big spike at the boundary. The size of the spike gets larger as the constraint gets more binding. The Bayesian appears to be roughly a truncated normal and does not spike at the boundary. Since these plots are sampling distributions and not posterior densities, the shapes are not surprising.

As is the case in the linear regression, imposing restrictions leads to biased estimates but improved variance. In order to evaluate the cost associated with the introduction of bias and the benefit of the lower standard deviation, we use the mean squared error criterion. Table 2.4 reports the mean square error for the estimates for both methods for the sample size of 100, and Table 2.7 for the corresponding values for a smaller sample size. Constrained maximum likelihood estimation leads to improvement

over the unconstrained estimation except for the last two cases where the constraints are most incorrect.

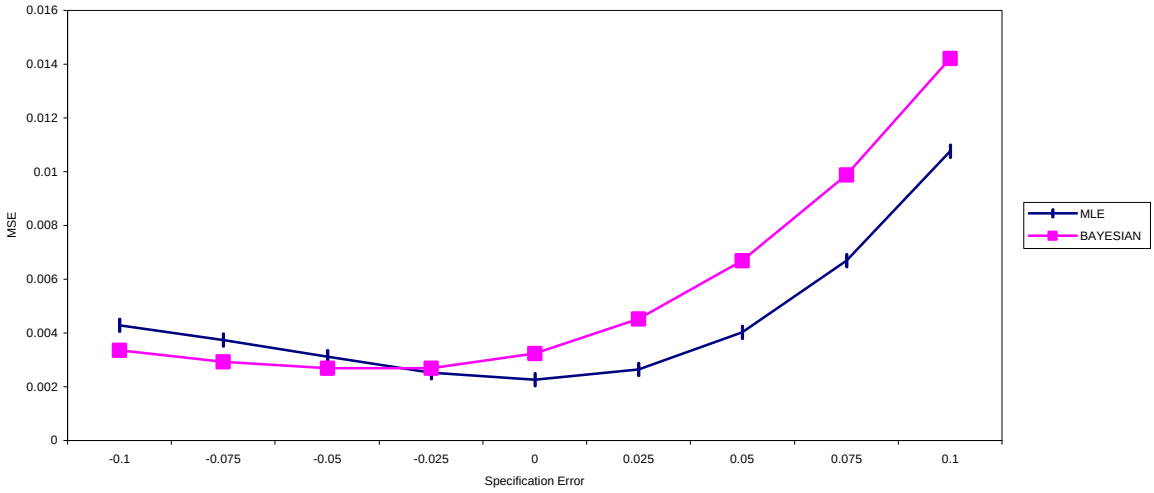


Figure 2.6: Mean Squared Error (MSE) Comparison for Beta3 (N=100)

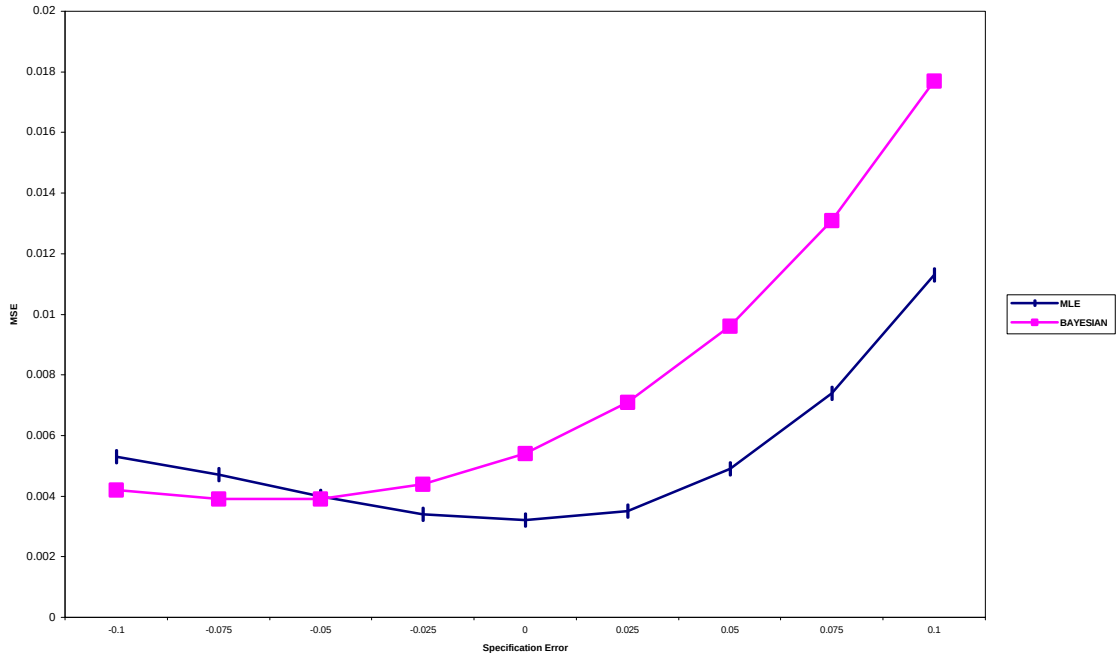


Figure 2.7: Mean Squared Error (MSE) Comparison for Beta3 (N=50)

In Figure 2.6, it can be seen that the MSE for the MLE hits its lowest point when the constraint is at the true parameter values of 0 and the specification error is zero. The MSEs for the other coefficients are not responsive to the specification error in b_3 due to the independence of the regressors. The only change in the trend that arose from a smaller sample size is with the bias of b_2. When the sample size is small, the MLE estimator for b_2 produces a larger bias than its Bayesian counterpart. The increase in the sample size benefits MLE faster in terms of bias. The variance, MSE and the bias for the other parameter estimators behave similarly in both of the experiments.

The risk behavior of the Bayesian estimator for b_3 is similar to that of MLE. It is lower when the inequality restriction is correct; however, it increases when it is incorrect at a higher rate than does the MLE. As can be seen from Table 2.4 and Figure 2.6, when the constraint is correct, the risk gain of the Bayesian estimator is much higher than that of the MLE, due to the higher efficiency. In those instances, the Bayesian does considerably better than MLE. In the Bayesian case, the minimum MSE is achieved before the point of zero specification error. Figure 2.7 represents the MSE of the Bayesian and ML methods when n=50. Despite the same overall trend, the minimum MSE is achieved earlier than it is with a larger sample size. Up to that point, nevertheless, the Bayesian estimator has a lower risk than its ML counterpart even in the small samples. The efficiency gain in the Bayesian case is large enough to justify the larger increase in bias. As in the case of MLE, the remaining parameter estimates are not affected by the restrictions on b_3 but indicate higher bias and MSE for every level of specification error.

The overall results indicate a consistent trend across different sample sizes. The efficiency gain of the Bayesian estimator under a correct inequality restriction is commendable. The bias generated by imposing a restriction on the parameter is more than compensated by the gain in efficiency, resulting in a decrease in the overall risk when the restrictions are correctly imposed. Given that the inequality restrictions will come from the economic theory, imposing these restrictions will in most cases improve the estimation process under the Bayesian technique.

2.4 Conclusion

The frequentist theory suggests that introducing inequality restrictions improves the estimation process to the extent that the restrictions are correct. The objective of the chapter is to compare the performance of ML and Bayesian probit estimators when different inequality restrictions are imposed. These restrictions cover different degrees of specification error. Comparisons are made on the basis of bias, variance and Mean Square Error (MSE). The experiment is conducted for the sample sizes of 50 and 100.

The results of the experiment showed that the MLE has lower MSE in the unconstrained case for both sample sizes and all parameter estimates. Higher sample size improves the estimation process for both techniques but faster for MLE than it is for Bayesian.

In the constrained case, an increase in the specification error increases the bias for both techniques. The Bayesian bias is higher than that of the MLE at each level of specification error. The difference does not change with the specification error. Variance of Bayesian estimator is lower in almost all cases. The difference decreases as the specification error increases. The variances of the estimators converge to each other

as the specification error gets more inaccurate. MSE for both techniques is lower when the restrictions are correct. It reaches its minimum level around the true value and as expected, increases as the restrictions gets more inaccurate. When the restrictions are correct, the Bayesian estimation has lower risk. The risk of Bayesian estimator increases as the specification error goes up at a faster rate than it does for MLE.

As an extension to this study, the bias, variance, MSE of the marginal effects of the probit coefficients, as well as the elasticities could be calculated in addition to the bias, variance and MSE of the coefficients. Griffiths, Hill and O'Donnell (2001) have shown that the results for the marginal probabilities and elasticities might yield results different that those obtained from the coefficients. Predictions using the constrained probabilities can also be investigated and compared under the techniques in question. The research can also be extended to other qualitative choice models such as the ordered or multinomial probit.

In addition, we can further improve the Bayesian estimation by imposing less conservative non-informative or even informative priors. That will result in even lower variances, thus a lower MSE. The change in the prior variance can delay the point of minimum for the Bayesian and can make Bayesian technique more tolerant to slightly wrong constraints.

2.5 References

Albert, and S. Chib (1993), "Bayesian Analysis of Binary and Polychhotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.

Amemiya, T., *Advanced Econometrics*, 1985, Harvard University Press, Cambridge, Massachusetts.

Casella, G. & E. George (1992), "Explaining the Gibbs sampler," *American Statistician,*46, 167-174.

Chen, M., and Q. Shao (1998), "Monte Carlo Methods for Bayesian Analysis of Constrained Parameter Problems," *Biometrika*, 85, 73-87.

Chib, S. (1992), "Bayes inference in the Tobit Regression Model," *Journal of Econometrics*, 51, 79-99.

Chib, S. and E. Greenberg (1996), "Markov Chain Monte Carlo simulation methods in Econometrics," *Economic Theory*, 12, 409-431.

Davis, W. W. (1978), "Bayesian Analysis of the Linear Model Subject to Linear Inequality Constraints," *Journal of American Statistical Association*, 73, 573-579.

Dhillon, U. S., Shilling, J. D., and C. F. Sirmans (1987), "Choosing Between Fixed and Adjustable Rate Mortgages," *Journal of Money, Credit, and Banking*, 19, 260-267.

Gelfand, A.E. and A.F.M. Smith (1990), "Sampling based approaches to calculating marginal densities," *Journal of American Statistical Association*, 85, 398-409.

Geman, S. and D. Geman (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12, 609-628.

Geweke, J. F. (1995), "Bayesian Inference for Linear Models Subject to Linear Inequality Constraints," Working Paper #552, University of Minnesota.

Geweke, J. F. (1991), "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints" in E. M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571-578. Fairfax: Interface Foundation of North America, Inc.

Green W. H., and T. Seaks (1989) "Use and Interpretation of the Lagrange Multipliers in Restricted Regression," University of North Carolina.

Griffiths, W. E., and D. Chotikapanich (1997), "A Comparison of the Finite Sample Properties of Maximum Likelihood and Bayesian Estimators for the Probit and the Tobit Models," Working Paper, University of New England, Australia.

Griffiths, W. E., and D. Chotikapanich (1996), "Bayesian Methodology for Imposing Inequality Constraints on a Linear Expenditure Function with Demographic Factors," *Australian Economic Papers*, forthcoming.

Griffiths, W. E., Hill, C. R., and C. J. O'Donnell (2001) "Including Prior Information in Probit Model Estimation"

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.

Judge, G. G. and T. Takayama (1966), "Inequality Restrictions in Regression Analysis," *Journal of the American Statistical Association*, 61, 166-181.

Judge, G. G., Griffiths, W. E., Hill R. C., Lütkepohl, H., and T. Lee, *The Theory and Practice of Econometrics*, 1985, John Wiley & Sons, New York, NY.

Judge, G. G., and Yancey, T. A. (1978) "Inequality Restricted Estimation Under Squared Error Loss," Working Paper, University of California, Berkley, CA.

Lovell M. C. and E. Prescott (1970), "Multiple Regression with Inequality Constraints: Pretesting, Bias, Hypothesis Testing and Efficiency," *Journal of the American Statistical Association*, 65, 913-925.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953), "Equations of the state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087-1092.

Tanner, M. A. and Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528-549.

Wei, G. C. G., and M. A. Tanner (1990), "Posterior computations for Censored Regression Data," *Journal of the American Statistical Association*, 85, 829-839.

Zellner, A. (1961), "Linear Regression with Inequality Constraints on the Coefficients," Mimeo #6109, The International Center for Management Science.

CHAPTER 3

ESTIMATION OF BINARY CHOICE MODEL UNDER ASYMMETRIC REGRESSORS

3.1 Introduction

In this chapter a Monte Carlo simulation is designed to assess the importance of the regressor distribution on the parameter estimates in finite samples. The data are generated according to the following model:

$$y_i^* = b_1 + b_2 x_{2i} + b_3 x_{3i} + e_i \qquad i = 1, 2, .., N$$

where x_2, x_3 are the regressors, e_i are the random errors that are assumed to be independently and identically distributed standard normal and the $b's$ are the unknown parameters. The dependent variable, y_i is created as

$$y_i = \begin{cases} 1 & if\ y_i^* > 0 \\ 0 & if\ y_i^* \leq 0 \end{cases}$$

Twelve different experimental designs are constructed by varying the true value of a parameter, the distribution of a regressor and the collinearity between the regressors for two sample sizes, yielding 24 total design points. Table 3.1 contains the description of each design. The following are the specifics of the experimental design:

(1) $b_1 = 0.9633,\ b_2 = -0.0973$

(2) $b_3 = 0.5$ or $b_3 = -0.5$

(3) x_2 and x_3 both $N(0,1)$ or x_2 and x_3 both c_4^2, or x_2 is $N(0,1)$ and x_3 is c_4^2.

In all cases c_4^2 variables are standardized to mean zero and variance 1.

(4) The standardized independent variables are multiplied with the root of a covariance matrix, where the variances are 1 and the covariances are 0 or 0.80, that results in varying collinearities.

(5) Sample sizes N = 50 and N = 100 are considered.

A total of 500 Monte Carlo samples are generated for each design point. For the MLE, the Newton-Raphson iterative algorithm is employed to maximize the probit log-likelihood function, using SAS/IML with the optimizer NLPNRA, and the means and standard deviations of the sampling distribution are calculated. For Bayesian estimates we use a non-informative prior of $b_i \sim N(0,100)$ as well as an informative prior of $b_i \sim N(0,4)$. Gibbs sampling is used as described in Chapter 2, to generate 1,100 iterations; 100 are discarded as part of the burn-in period. The posterior means, standard deviations and bias are calculated for each sample and the sampling distributions by averaging over 500 Monte Carlos for each design point. To convince ourselves of the convergence of the Gibbs algorithm, informal checks of changes in parameter values are done for varying iterations for some data sets. We will present only one since all the results were similar. For experimental design 2, the Gibbs was run for 11,000 iterations and 1,000 were discarded. The changes in the mean of the sampling distributions were insignificant. The change in the bias for b_1 was from 0.1347 to 0.1324, for b_2 from -0.0298 to -0.0302, and for b_3, from 0.1106 to 0.1093. Comparison of the MLE and the Bayesian method has been performed on the basis of bias, variance and MSE. Different numbers of iterations have been tried. The changes in the variance and MSE due to those different Gibbs iterations were also small. These small changes did not alter the results of the comparisons between the methods for the design points tested.

Table 3.1: Descriptions of Experimental Design

		Low correlation	Low correlation	High correlation	High correlation
		$b_3 = -0.5$	$b_3 = 0.5$	$b_3 = -0.5$	$b_3 = 0.5$
N=50	X2=X3= $N(0,1)$	**1**	2	**3**	4
	X2= $N(0,1)$, X3= c_4^2	**5**	6	7	8
	X2=X3= c_4^2	**9**	10	11	12
N=100	X2=X3= $N(0,1)$	**13**	14	15	16
	X2= $N(0,1)$, X3= c_4^2	17	18	19	20
	X2=X3= c_4^2	21	22	23	24

3.2 Experimental Design 1

The results for the first experimental design are presented in Tables 3.2, 3.3, 3.4, and 3.5 for b_2 and b_3, and Tables 3.6-3.8 for the intercept.

Table 3.2: Descriptive Statistics for the Regressors Experimental Design 1 and 2

	X2	X3
Mean	0.1103	-0.0026
Variance	0.6277	1.4031
Skewness	-0.878	0.0865
Kurtosis	0.1523	0.4641
Correlation	-0.1535	-0.1535
# of samples	500	500

For this design the sample size is N=50 and regressors are both standard normal with no collinearity. Figure B.1 illustrates the histogram for these regressors. As can be seen from the descriptive statistics, as well as the histogram in Figure B.1, both of the regressors are centered very close to 0 and have slight skewness but are very close to being symmetric. The skewness is the measure of the asymmetry of a distribution. A symmetric distribution is one where $f(m-x) = f(m+x)$, whereas an asymmetric

distribution has a long tail in positive or negative direction that results in a positive or a negative skew, respectively. The measure of skewness is $\frac{m^3}{s^3}$. Another density characteristic is kurtosis which measures the thickness of the tail of a distribution, and the excess kurtosis value is estimated as $\frac{m^4}{s^4} - 3$. Both values are zero for the normal distribution.

In the first experimental design, both distributions of the regressors are consistent with a standard normal density function. Based on these regressors, MSE, bias and variance of the MLE and Bayesian point estimates are reported in Tables 3.3, 3.4, and 3.5 respectively.

It can be seen that both the ML and Bayesian estimates have an upward bias in absolute value with the ML estimates for all parameters having slightly smaller bias than their Bayesian counterparts. Also in Figure B.1 are the histograms for the sampling distribution for the ML and Bayesian estimates for b_2 and b_3. The sampling distribution of b_2 for both estimation techniques is symmetric and very similar, with a slightly larger skew for the Bayesian estimate. The posterior distribution of b_3 is more skewed in both cases, to the left, with a more pronounced skew for the Bayesian estimate.

The variances of the estimates are comparable for all parameters. Due to slightly larger extremes in the Bayesian case, the variance is somewhat larger. The non-informative prior imposed on the parameters of the Bayesian model can be changed to more informative priors that will lower the bias as well as the variance of the estimators.

The MSE of the parameters are calculated as:

$$MSE = bias^2 + variance$$

Due to the lower bias and variance of the ML estimates for all parameters, the MLE outperforms the Bayesian in terms of MSE, as well. MLE produces MSEs that are 80 percent of the Bayesian estimate for the location parameter, 90 percent for the b_2, and 85 percent for the b_3. These results can be seen in Tables 3.3, 3.11 and 3.6.

For this specific design, the results are intuitive. MLE is expected to do well when the regressors are symmetric, the *y*s are balanced and the sample size is not too small. On the other hand, the prior variance, which is 100 for Bayesian parameters, is very conservative. We almost always have a better idea about the distribution of the parameters and the variance is never expected to be 100. The MSE, bias, and variance results of the more informative prior for Bayesian parameters are also reported in Tables 3.3, 3.4 and 3.5 respectively. Figure B.1 indicates no drastic changes in the sampling distribution of the parameters when the informative prior is used. The only change is the elimination of very unlikely values. As a result we observe improvement of the bias, variance, and the MSE of informative estimation over the non-informative for all parameters. Figure B.1 also illustrates the sampling distribution of b_2 and b_3 using the prior information of more realistic values for prior variance. The incorporation of the informative prior also improves the estimation process over the MLE. The improvement of the variance especially contributes to the decline in the MSE for the Bayesian estimation.

Table 3.3 : MSE of Estimators with Uncorrelated Regressors. b_3 =-0.5, N=50						
Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design (Design #)	b_2	b_3	b_2	b_3	b_2	b_3
X2=X3= $N(0,1)$ (1)	0.1220 (0.0111)	0.0751 (0.0092)	0.1351 (0.0138)	0.0885 (0.0125)	0.1145 (0.0090)	0.0699 (0.0060)
X2= $N(0,1)$, X3= c_4^2 (5)	0.0879 (0.0127)	0.0782 (0.0073)	0.0945 (0.0121)	0.0876 (0.0082)	0.0775 (0.0081)	0.0752 (0.0066)
X2=X3= c_4^2 (9)	0.0731 (0.0053)	0.0705 (0.0055)	0.0816 (0.0062)	0.0796 (0.0065)	0.0741 (0.0054)	0.0689 (0.0052)
b_3 =0.5, N=50						
X2=X3= $N(0,1)$ (2)	0.1281 (0.0107)	0.0918 (0.0088)	0.1413 (0.0123)	0.1047 (0.0100)	0.1206 (0.0091)	0.0892 (0.0081)
X2= $N(0,1)$, X3= c_4^2 (6)	0.0599 (0.0044)	0.2173 (0.0427)	0.0645 (0.0049)	0.2315 (0.0350)	0.0581 (0.0042)	0.1520 (0.0146)
X2=X3= c_4^2 (10)	0.0732 (0.0083)	0.2284 (0.0471)	0.0803 (0.0091)	0.2431 (0.0351)	0.0691 (0.0062)	0.1592 (0.0165)
b_3 =-0.5, N=100						
X2=X3= $N(0,1)$ (13)	0.0437 (0.0031)	0.0253 (0.0019)	0.0457 (0.0033)	0.0268 (0.0020)	0.0433 (0.0031)	0.0258 (0.0020)
X2= $N(0,1)$, X3= c_4^2 (17)	0.0336 (0.0031)	0.0301 (0.0023)	0.0351 (0.0025)	0.0321 (0.0025)	0.0338 (0.0023)	0.0306 (0.0023)
X2=X3= c_4^2 (21)	0.0291 (0.0019)	0.0293 (0.0024)	0.0308 (0.0020)	0.0314 (0.0026)	0.0298 (0.0020)	0.0295 (0.0023)
b_3 =0.5, N=100						
X2=X3= $N(0,1)$ (14)	0.0353 (0.0024)	0.0291 (0.0023)	0.0366 (0.0025)	0.0312 (0.0025)	0.0355 (0.0024)	0.0291 (0.0022)
X2= $N(0,1)$, X3= c_4^2 (18)	0.0332 (0.0020)	0.0613 (0.0054)	0.0346 (0.0021)	0.0685 (0.0062)	0.0328 (0.0020)	0.0591 (0.0051)
X2=X3= c_4^2 (22)	0.0320 (0.0022)	0.0592 (0.0060)	0.0336 (0.0023)	0.0666 (0.0070)	0.0325 (0.0022)	0.0572 (0.0056)

Table 3.4 : Bias of Estimators with Uncorrelated Regressors. b_3 =-0.5, N=50						
Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design (Design #)	b_2	b_3	b_2	b_3	b_2	b_3
X2=X3= $N(0,1)$ (1)	-0.0183	-0.0670	-0.0207	-0.0883	-0.0111	-0.0688
X2= $N(0,1)$, X3= c_4^2 (5)	-0.0530	-0.0751	-0.0631	-0.0920	-0.0476	-0.0725
X2=X3= c_4^2 (9)	-0.0094	-0.0658	-0.0074	-0.0798	-0.0036	-0.0633
b_3 =0.5, N=50						
X2=X3= $N(0,1)$ (2)	-0.0256	0.0875	-0.0298	0.1106	-0.0242	0.0905
X2= $N(0,1)$, X3= c_4^2 (6)	-0.0424	0.1375	-0.0493	0.1732	-0.0392	0.1244
X2=X3= c_4^2 (10)	-0.0040	0.1430	-0.0003	0.1818	0.0020	0.1327
b_3 =-0.5, N=100						
X2=X3= $N(0,1)$ (13)	-0.0074	-0.0229	-0.0094	-0.0326	-0.0067	-0.0272
X2= $N(0,1)$, X3= c_4^2 (17)	-0.0100	-0.0446	-0.0118	-0.0517	-0.0109	-0.0462
X2=X3= c_4^2 (21)	-0.0079	-0.0411	-0.0073	-0.0485	-0.0064	-0.0415
b_3 =0.5, N=100						
X2=X3= $N(0,1)$ (14)	-0.0112	0.0346	-0.0127	0.0432	-0.0113	0.0361
X2= $N(0,1)$, X3= c_4^2 (18)	-0.0132	0.0586	-0.0156	0.0780	-0.0134	0.0630
X2=X3= c_4^2 (22)	0.0146	0.0554	0.0158	0.0734	0.0167	0.0593

Table 3.5
Variance of Estimators with Uncorrelated Regressors.
b_3 =-0.5, N=50

Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design (Design #)	b_2	b_3	b_2	b_3	b_2	b_3
X2=X3= $N(0,1)$ (1)	0.1216	0.0707	0.1347	0.0807	0.1144	0.0651
X2= $N(0,1)$, X3= c_4^2 (5)	0.0851	0.0726	0.0905	0.0792	0.0753	0.0700
X2=X3= c_4^2 (9)	0.0730	0.0661	0.0816	0.0702	0.0741	0.0649
b_3 =0.5, N=50						
X2=X3= $N(0,1)$ (2)	0.1275	0.0842	0.1404	0.0925	0.1200	0.0810
X2= $N(0,1)$, X3= c_4^2 (6)	0.0581	0.1984	0.0621	0.2015	0.0565	0.1365
X2=X3= c_4^2 (10)	0.0732	0.2080	0.0803	0.2101	0.0691	0.1416
b_3 =-0.5, N=100						
X2=X3= $N(0,1)$ (13)	0.0436	0.0248	0.0456	0.0258	0.0432	0.0251
X2= $N(0,1)$, X3= c_4^2 (17)	0.0335	0.0282	0.0349	0.0294	0.0336	0.0284
X2=X3= c_4^2 (21)	0.0290	0.0276	0.0308	0.0291	0.0297	0.0278
b_3 =0.5, N=100						
X2=X3= $N(0,1)$ (14)	0.0352	0.0279	0.0364	0.0294	0.0354	0.0278
X2= $N(0,1)$, X3= c_4^2 (18)	0.0330	0.0579	0.0344	0.0624	0.0326	0.0551
X2=X3= c_4^2 (22)	0.0317	0.0562	0.0334	0.0612	0.0322	0.0537

Table 3.6
MSE of b_1 Estimators
b_3 =-0.5, N=50

Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.1232 (1)	0.1055 (3)	0.1512 (1)	0.1284 (3)	0.1089 (1)	0.0976 (3)
X2= $N(0,1)$, X3= c_4^2	0.1284 (5)	0.1504 (7)	0.1463 (5)	0.2005 (7)	0.1074 (5)	0.1209 (7)
X2=X3= c_4^2	0.0843 (9)	0.1283 (11)	0.0983 (9)	0.1397 (11)	0.0805 (9)	0.0928 (11)
b_3 =0.5, N=50						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.1089 (2)	0.0955 (4)	0.1268 (2)	0.1088 (4)	0.1009 (2)	0.0878 (4)
X2= $N(0,1)$, X3= c_4^2	0.1551 (6)	0.4002 (8)	0.1709 (6)	0.2083 (8)	0.1131 (6)	0.0973 (8)
X2=X3= c_4^2	0.1550 (10)	0.1680 (12)	0.1696 (10)	0.1771 (12)	0.1068 (10)	0.0993 (12)
b_3 =-0.5, N=100						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0361 (13)	0.0380 (15)	0.0393 (13)	0.0408 (15)	0.0364 (13)	0.0382 (15)
X2= $N(0,1)$, X3= c_4^2	0.0327 (17)	0.0360 (19)	0.0348 (17)	0.0391 (19)	0.0328 (17)	0.0357 (19)
X2=X3= c_4^2	0.0318 (21)	0.0344 (23)	0.0339 (21)	0.0372 (23)	0.0318 (21)	0.0342 (23)
b_3 =0.5, N=100						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0371 (14)	0.0355 (16)	0.0399 (14)	0.0381 (16)	0.0370 (14)	0.0352 (16)
X2= $N(0,1)$, X3= c_4^2	0.0385 (18)	0.0358 (20)	0.0440 (18)	0.0390 (20)	0.0381 (18)	0.0349 (20)
X2=X3= c_4^2	0.0396 (22)	0.0369 (24)	0.0452 (22)	0.0414 (24)	0.0396 (22)	0.0371 (24)

Table 3.7
Bias of b_1 Estimators
b_3 =-0.5, N=50

Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.1125 (1)	0.1049 (3)	0.1411 (1)	0.1336 (3)	0.1092 (1)	0.1012 (3)
X2= $N(0,1)$, X3= c_4^2	0.1077 (5)	0.1255 (7)	0.1349 (5)	0.1613 (7)	0.1001 (5)	0.1132 (7)
X2=X3= c_4^2	0.0811 (9)	0.1035 (11)	0.1014 (9)	0.1253 (11)	0.0765 (9)	0.0895 (11)
b_3 =0.5, N=50						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.1063 (2)	0.0972 (4)	0.1347 (2)	0.1221 (4)	0.1037 (2)	0.0926 (4)
X2= $N(0,1)$, X3= c_4^2	0.1445 (6)	0.1555 (8)	0.1796 (6)	0.1773 (8)	0.1353 (6)	0.0848 (8)
X2=X3= c_4^2	0.1355 (10)	0.1458 (12)	0.1700 (10)	0.1792 (12)	0.1252 (10)	0.1274 (12)
b_3 =-0.5, N=100						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0437 (13)	0.0468 (15)	0.0563 (13)	0.0582 (15)	0.0472 (13)	0.0483 (15)
X2= $N(0,1)$, X3= c_4^2	0.0406 (17)	0.0436 (19)	0.0497 (17)	0.0534 (19)	0.0408 (17)	0.0443 (19)
X2=X3= c_4^2	0.0374 (21)	0.0424 (23)	0.0462 (21)	0.0516 (23)	0.0376 (21)	0.0420 (23)
b_3 =0.5, N=100						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0443 (14)	0.0400 (16)	0.0554 (14)	0.0501 (16)	0.0450 (14)	0.0397 (16)
X2= $N(0,1)$, X3= c_4^2	0.0558 (18)	0.0515 (20)	0.0720 (18)	0.0644 (20)	0.0590 (18)	0.0530 (20)
X2=X3= c_4^2	0.0553 (22)	0.0576 (24)	0.0712 (22)	0.0734 (24)	0.0579 (22)	0.0614 (24)

Table 3.8
Variance of b_1 Estimators
b_3 =-0.5, N=50

Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.1106 (1)	0.0945 (3)	0.1312 (1)	0.1106 (3)	0.0970 (1)	0.0873 (3)
X2= $N(0,1)$, X3= c_4^2	0.1168 (5)	0.1346 (7)	0.1281 (5)	0.1745 (7)	0.0974 (5)	0.1081 (7)
X2=X3= c_4^2	0.0778 (9)	0.1176 (11)	0.0880 (9)	0.1240 (11)	0.0746 (9)	0.0848 (11)
b_3 =0.5, N=50						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0976 (2)	0.0861 (4)	0.1087 (2)	0.0939 (4)	0.0902 (2)	0.0792 (4)
X2= $N(0,1)$, X3= c_4^2	0.1343 (6)	0.3760 (8)	0.1387 (6)	0.1773 (8)	0.0948 (6)	0.0848 (8)
X2=X3= c_4^2	0.1366 (10)	0.1468 (12)	0.1407 (10)	0.1450 (12)	0.0911 (10)	0.0831 (12)
b_3 =-0.5, N=100						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0342 (13)	0.0358 (15)	0.0361 (13)	0.0374 (15)	0.0342 (13)	0.0359 (15)
X2= $N(0,1)$, X3= c_4^2	0.0310 (17)	0.0341 (19)	0.0323 (17)	0.0362 (19)	0.0311 (17)	0.0337 (19)
X2=X3= c_4^2	0.0304 (21)	0.0424 (23)	0.0317 (21)	0.0516 (23)	0.0304 (21)	0.0420 (23)
b_3 =0.5, N=100						
X Design	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)	Uncorrelated Regressors (Design #)	Correlated Regressors (Design #)
X2=X3= $N(0,1)$	0.0351 (14)	0.0339 (16)	0.0368 (14)	0.0356 (16)	0.0350 (14)	0.0337 (16)
X2= $N(0,1)$, X3= c_4^2	0.0354 (18)	0.0331 (20)	0.0388 (18)	0.0348 (20)	0.0347 (18)	0.0304 (20)
X2=X3= c_4^2	0.0366 (22)	0.0336 (24)	0.0401 (22)	0.0360 (24)	0.0362 (22)	0.0334 (24)

Table 3.9: Asymptotic Variance of Selected Design Points				
Design Point	N=50		N=100	
	Variance of b_2	Variance of b_3	Variance of b_2	Variance of b_3
5 vs. 17	0.0443	0.0516	0.0255	0.0205
9 vs. 21	0.0473	0.0466	0.0229	0.0201
11 vs. 23	0.1534	0.1624	0.0573	0.0632

3.3 Experimental Design 3

In experimental design 3, the only modification is to introduce correlation between the regressors.

Table 3.10: Descriptive Statistics for the Regressors Experimental Design 3 and 4

	X2	X3
Mean	0.1103	0.0866
Variance	0.6277	0.7685
Skewness	-0.878	0.3392
Kurtosis	0.1523	-0.1371
Correlation	0.5985	0.5985
# of samples	500	500

Instead of the –15 percent correlation of the first design, design 3 has 60 percent correlation between the regressors which is expected to delay the asymptotic properties of MLE. As can be seen from Table 3.14, the regressors are still representative of the normal density but X3 has been multiplied with a matrix to simulate collinearity between X2 and X3. Figure B.3 depicts the distribution of the regressors. Table 3.15 illustrates the changes in the X3, which is still centered around zero but much less symmetric.

As shown in Tables 3.7 and 3.12, respectively, the upward bias of the estimates still exists with smaller absolute values for b_1 and b_2. The introduction of high correlation increases the bias of estimation for all three processes for b_3. In finite samples, Bayesian estimator reacts similar to the MLE to the introduction of collinearity.

Table 3.11: MSE of Estimators with Correlated Regressors. b_3 =-0.5, N=50						
Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design (Design #)	b_2	b_3	b_2	b_3	b_2	b_3
X2=X3= $N(0,1)$ (3)	0.1514 (0.0095)	0.1513 (0.0129)	0.1643 (0.0105)	0.1702 (0.0150)	0.1375 (0.0083)	0.1383 (0.0111)
X2= $N(0,1)$, X3= c_4^2 (7)	0.1707 (0.0130)	0.1678 (0.0118)	0.1901 (0.0157)	0.1831 (0.0132)	0.1535 (0.0108)	0.1489 (0.0101)
X2=X3= c_4^2 (11)	0.1599 (0.0131)	0.2271 (0.0297)	0.1742 (0.0146)	0.2430 (0.0246)	0.1479 (0.0116)	0.1777 (0.0123)
b_3 =0.5, N=50						
X2=X3= $N(0,1)$ (4)	0.1592 (0.0112)	0.1880 (0.0150)	0.1688 (0.0118)	0.2041 (0.0160)	0.1452 (0.0097)	0.1636 (0.0119)
X2= $N(0,1)$, X3= c_4^2 (8)	0.6308 (0.2603)	1.0133 (0.4896)	0.4588 (0.0660)	0.6632 (0.1089)	0.2191 (0.0159)	0.2779 (0.0224)
X2=X3= c_4^2 (12)	0.3997 (0.0585)	0.4936 (0.1150)	0.4030 (0.0439)	0.5181 (0.0971)	0.2386 (0.0171)	0.2515 (0.0231)
b_3 =-0.5, N=100						
X2=X3= $N(0,1)$ (15)	0.0715 (0.0048)	0.0604 (0.0043)	0.0737 (0.0050)	0.0626 (0.0045)	0.0698 (0.0046)	0.0587 (0.0042)
X2= $N(0,1)$, X3= c_4^2 (19)	0.0852 (0.0064)	0.0690 (0.0051)	0.0889 (0.0066)	0.0721 (0.0058)	0.0831 (0.0063)	0.0664 (0.0048)
X2=X3= c_4^2 (23)	0.0589 (0.0043)	0.0683 (0.0046)	0.0610 (0.0045)	0.0731 (0.0051)	0.0578 (0.0042)	0.0658 (0.0044)
b_3 =0.5, N=100						
X2=X3= $N(0,1)$ (16)	0.0676 (0.0043)	0.0713 (0.0054)	0.0697 (0.0045)	0.0754 (0.0058)	0.0652 (0.0042)	0.0680 (0.0051)
X2= $N(0,1)$, X3= c_4^2 (20)	0.1245 (0.0091)	0.1264 (0.0117)	0.1291 (0.0093)	0.1339 (0.0119)	0.1125 (0.0075)	0.1123 (0.0089)
X2=X3= c_4^2 (24)	0.1075 (0.0085)	0.1151 (0.0114)	0.1138 (0.0089)	0.1245 (0.0123)	0.0988 (0.0071)	0.1019 (0.0087)

Table 3.12: Bias of Estimators with Correlated Regressors. b_3 =-0.5, N=50						
Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design (Design #)	b_2	b_3	b_2	b_3	b_2	b_3
X2=X3= $N(0,1)$ (3)	-0.0020	-0.0838	-0.0066	-0.1042	-0.0075	-0.0758
X2= $N(0,1)$, X3= c_4^2 (7)	-0.0782	-0.0378	-0.0928	-0.0529	-0.0838	-0.0241
X2=X3= c_4^2 (11)	-0.0495	-0.0602	-0.0571	-0.0746	-0.0632	-0.0397
b_3 =0.5, N=50						
X2=X3= $N(0,1)$ (4)	-0.0244	0.0836	-0.0302	0.1071	-0.0083	0.0695
X2= $N(0,1)$, X3= c_4^2 (8)	-0.1565	0.2254	-0.1723	0.2552	-0.0696	0.1258
X2=X3= c_4^2 (12)	-0.0489	0.1713	-0.0527	0.2136	0.0042	0.1193
b_3 =-0.5, N=100						
X2=X3= $N(0,1)$ (15)	-0.0128	-0.0247	-0.0140	-0.0336	-0.0159	-0.0240
X2= $N(0,1)$, X3= c_4^2 (19)	-0.0190	-0.0269	-0.0256	-0.0318	-0.0280	-0.0238
X2=X3= c_4^2 (23)	0.0022	-0.0474	-0.0012	-0.0551	-0.0044	-0.0442
b_3 =0.5, N=100						
X2=X3= $N(0,1)$ (16)	-0.0337	0.0598	-0.0350	0.0688	-0.0247	0.0541
X2= $N(0,1)$, X3= c_4^2 (20)	-0.0144	0.0494	-0.0242	0.0691	-0.0060	0.0449
X2=X3= c_4^2 (24)	0.0191	0.0591	0.0162	0.0794	0.0306	0.0559

Table 3.13
Variance of Estimators with Correlated Regressors.
b_3 =-0.5, N=50

Estimators	MLE		Noninformative Bayesian		Informative Bayesian	
X Design (Design #)	b_2	b_3	b_2	b_3	b_2	b_3
X2=X3= $N(0,1)$ (3)	0.1514	0.1443	0.1642	0.1593	0.1374	0.1325
X2= $N(0,1)$, X3= c_4^2 (7)	0.1645	0.1664	0.1815	0.1803	0.1465	0.1483
X2=X3= c_4^2 (11)	0.1575	0.2235	0.1709	0.2375	0.1440	0.1761
b_3 =0.5, N=50						
X2=X3= $N(0,1)$ (4)	0.1586	0.1810	0.1679	0.1927	0.1451	0.1588
X2= $N(0,1)$, X3= c_4^2 (8)	0.6063	0.9626	0.4291	0.5981	0.2142	0.2621
X2=X3= c_4^2 (12)	0.3974	0.4643	0.4002	0.4725	0.2386	0.2372
b_3 =-0.5, N=100						
X2=X3= $N(0,1)$ (15)	0.0714	0.0598	0.0735	0.0615	0.0696	0.0581
X2= $N(0,1)$, X3= c_4^2 (19)	0.0848	0.0683	0.0883	0.0711	0.0824	0.0658
X2=X3= c_4^2 (23)	0.0589	0.0660	0.0610	0.0701	0.0578	0.0638
b_3 =0.5, N=100						
X2=X3= $N(0,1)$ (16)	0.0664	0.0677	0.0684	0.0707	0.0646	0.0651
X2= $N(0,1)$, X3= c_4^2 (20)	0.1243	0.1240	0.1285	0.1291	0.0297	0.0278
X2=X3= c_4^2 (24)	0.1071	0.1116	0.1135	0.1182	0.0979	0.0988

Figure B.3 contains the sampling distributions of MLE, Bayesian and Bayesian with smaller prior variance. Comparison of these plots with the plots of the baseline case of design 1 confirms the numerical findings in Tables 3.11, 3.12, and 3.13 for MSE, bias, and variance respectively. Figure B.1 indicates a larger tail to the left for b_3 after the introduction of collinearity. That explains the increased variance and MSE for this parameter estimate. The impact is smaller in b_2. The same is true for the Bayesian parameter estimates for both with and without prior information as can be observed in Figure B.3.

Despite the smaller extreme values and less skewed distributions, the variances of b_2 and b_3 have gone up for all estimation processes. On the other hand, the variance of b_1 declines about 10 percent. However, the decline of the variance of b_1 (Table 3.9) is not as significant as the increase in the variance of b_2 and b_3 (Table 3.13). b_1, with its largest absolute true value, has a lower variance than the estimates of the parameters with smaller true values. Therefore, we can conclude that introduction of collinearity increases the overall variance of the estimation.

The MSE of b_1 insignificantly decreases, while increasing for the remaining parameter estimates as shown in Tables 3.6, and 3.11. The decline in the bias of b_2, is more than outweighed by the increase in the variance. The increase in MSE for b_3 is almost doubled for all methods after the introduction of collinearity. The MSE for Bayesian method remains higher than that of MLE. However, lowering the prior variance lowers the MSE of the Bayesian method below the MLE, as was the case with the first experimental design. Overall, the introduction of collinearity only changes the levels of the risk factors, but not the behavior patterns of the different methods.

3.4 Experimental Design 5

The purpose of experimental design 5 is to compare the performance of the estimation process when one of the regressors is skewed.

Table 3.14: Descriptive Statistics for the Regressors Experimental Design 5 and 6

	X2	X3
Mean	0.1776	0.0824
Variance	1.0438	1.0377
Skewness	-0.3062	0.7966
Kurtosis	-0.9233	-0.0494
Correlation	0.0671	0.0671
# of samples	500	500

In this design, X2 is standard normal and X3 has a standardized c_4^2 distribution. Figure B.5 indicates the distribution of X2 and X3. As can be seen from the figure, X3 has a c^2 distribution that explains the skew to the right. X2 is skewed to the left more than expected, due to the small sample size and randomness of the sample. Table 3.15 captures the descriptive statistics for the regressors. Figure B.5 presents the sampling distributions of the MLE, Bayesian and informative Bayesian parameters. Despite the skewness of the Xs, the sampling distributions of the parameters are not very different than they were in design 1.

One of the explanations of the minor difference is the bias, as can be seen in Table 3.4 for b_2 and b_3, and Table 3.7 for b_3. The bias for b_2 and b_3 improves in all methods, while the bias of b_1 increases slightly. However, even when one of the regressors is skewed, the Bayesian has the highest bias, followed by MLE. Prior information improves the bias in this case as well.

The behavior of variance is almost always favorable in the non-informative Bayesian method. The variance goes down for all estimates in the case of the non-informative prior. For

the MLE and informative Bayesian cases, it goes down only for b_2. With the non-informative Bayesian, the distribution is so spread due to the high prior variance that it is intuitive for the variance to go down when the regressors skew.

The MSE results are hard to interpret due to mixed bias and variance results. However, the overall results indicate that the Bayesian better handles the skewness in the regressor than does the MLE.

3.5 Experimental Design 9

Design 9 illustrates regressors that are both skewed. This skewness is obtained by sampling from two independent c_4^2 distributions and standardizing them.

Table 3.15: Descriptive Statistics for the Regressors Experimental Design 9 and 10

	X2	X3
Mean	0.0863	0.0824
Variance	0.9143	1.0377
Skewness	1.2697	0.7966
Kurtosis	1.2479	-0.0494
Correlation	-0.0661	-0.0661
# of samples	500	500

As a result we obtained two regressors with long tails to the right. X3 in design 9 is identical to the X3 in design 5. Table 3.16 contains the description for the Xs and illustrates the larger skewness values for both of the regressors. Figure B.9, depicts the skewness distributions for the regressors. Sampling distribution for b_2 and b_3 for different methods are in Figure B.9. These distributions are very similar for different methods and also resemble closely the sampling distributions of the design 1 parameters.

The surprising finding is that the bias, variance and MSE for all parameter estimates in all three methods improve when both regressors are skewed. The improvement is both over the bias of the baseline design 1 estimates, and also over the estimates of design 5, where only one

regressor is skewed. The improvement in the MLE is more pronounced, with a 70 percent drop in the MSE for b_1 and b_2 (Tables 6 and 3, respectively), compared to about 50 percent drop in the Bayesian case for the same parameters. The informative Bayesian has the lowest gain of all.

As was mentioned in Chapter 2, with a non-informative prior on b_3, the MLE is essentially the posterior mode. The Bayesian estimator under the quadratic loss function is the posterior mean. Because the mean lies to the left of the mode when skewness is to the left, the Bayesian method yields smaller estimates for b_3; whereas, when the distribution is symmetric, because the mode is equal to the mean, estimates are expected to mimic each other. It can be seen that MLE estimation improves with skewness as long as the skewness of the regressors is in the same direction. The MLE improves more than both the informative and non-informative Bayesian estimators. This is surprising, as both the non-informative and informative Bayesian estimators did better when only one regressor was skewed.

3.6 Experimental Design 13

This design has a sample size N=100 but is identical to the first design in all other ways. Table 3.18 has the descriptions of the Xs and we can see from Figure B.13, that the distributions very closely resemble the normal density.

Table 3.16: Descriptive Statistics for the Regressors Experimental Design 13 and 14

	X2	X3
Mean	0.0317	-0.0320
Variance	0.7342	1.3127
Skewness	0.0232	-0.0541
Kurtosis	-0.1052	-0.0433
Correlation	-0.0345	-0.0345
# of samples	500	500

Table 3.9 illustrates the asymptotic variance of the MLE estimator for b_2 and b_3 for selected design points. Comparison of these values to the empirical variances presented in Tables 3.5 and 3.13 indicates that when the sample size is 100, the MLE achieves its asymptotic properties for all the design points presented, whereas for sample size 50, the empirical variances are far larger than their asymptotic counterparts.

All methods for all parameter estimates improve with increased sample size, with MLE having the largest benefit and informative Bayesian the smallest. The risk results of the MLE and informative Bayesian are almost identical. All these figures are larger in the non-informative Bayesian case. However, observation of the Figure B.13 does not represent big differences in the sampling distribution of the parameters among the different methods. However, Tables 3.3, 3.4, and 3.5 indicate small numerical differences mentioned above.

Design 13 needs to be compared to designs 15, 17, and 21 to validate the impacts of changes in correlation, one regressor distribution, and two regressor distribution, respectively. It is beneficial to see if the relationships that hold for sample size N=50 hold for N=100.

3.7 Experimental Designs 15, 17 and 21

These designs are used to make comparisons to design 13. These comparisons will be analogous to the comparisons of designs 3, 5, and 9 to design 1, in the context of a small sample.

Table 3.17: Descriptive Statistics for the Regressors Experimental Design 15 and 16

	X2	X3
Mean	0.0317	0.0031
Variance	0.7341	0.9090
Skewness	0.0232	-0.2291
Kurtosis	-0.1052	0.1249
Correlation	0.6934	0.6934
# of samples	500	500

Table 3.18: Descriptive Statistics for the Regressors Experimental Design 17 and 18

	X2	X3
Mean	-0.0148	0.0844
Variance	0.8979	1.0467
Skewness	0.0919	1.1761
Kurtosis	0.2239	1.4971
Correlation	0.1302	0.1302
# of samples	500	500

Table 3.19: Descriptive Statistics for the Regressors Experimental Design 21 and 22

	X2	X3
Mean	-0.0200	0.0844
Variance	0.8921	1.0467
Skewness	1.4204	1.1761
Kurtosis	2.8502	1.4971
Correlation	-0.0738	-0.0738
# of samples	500	500

Design point 15 has a larger correlation between the Xs, as opposed to the low correlation of design 13. The only difference between designs 17 and 13 is the distribution of X3, and the difference between designs 13 and 21 is the distribution of X2 and X3. The purpose of these comparisons is to observe the above comparisons when the sample size gets larger.

The statistics of the Xs and the MSE, bias and variances of the estimation results of design 15 are illustrated in Tables 3.17, 3.11, 3.12, and 3.13 for b_2 and b_3 and Tables 3.6-3.8 for the intercept, respectively. Observation of the Xs indicates lower extreme values and less skewness. Figure B.15 indicates distributions close to the distribution of a normal density. It can be seen from Tables 3.11, 3.12 and 3.13 versus Tables 3.3, 3.4, and 3.5 that increasing the sample size from 50 to 100 lowers all the risk measures for all techniques except for the bias of β_2 for the informative prior. However, the MSE even in this case, went down due to lower variance. Tables 3.3, 3.11 and 3.6 indicate that the informative Bayesian outperforms the MLE

except for the estimation of the location parameter. Figure B.15 illustrates the risk gains as a result of increasing sample size on the sampling distributions. MLE seems to have the biggest gain although informative Bayesian does better in some instances. Despite improvements in the non-informative Bayesian case, the improvements are not large enough. In all the cases, MLE outperforms the non-informative Bayesian. Comparison of the risk measures between the MLE and the informative Bayesian are very mixed. For the location parameter, the MLE has a lower bias, variance, and thus MSE (Tables 3.7, 3.8, 3.6, respectively). The lower bias of MLE in the case of β_2 is outweighed by the increased variance, yielding a larger MSE. In all the other instances, the informative-Bayesian does better than the MLE. We can also observe the adverse effects of introducing correlation to the design matrix by comparison of designs 13 and 15. It shows increasing risk factors, except for the bias of informative β_3.

Design point 17 introduces a skewed regressor in the large sample size. Compared to the small sample case, design 17 has lower bias overall. The Bayesian still has the highest risk, which decreased with the introduction of prior. However, MLE did better than the informative Bayesian in many instances, as well. This result might the artifact of lower skewness of the Xs as presented in Table 3.3 compared to its small sample counterpart. Due to the large sample size all the variances go down; however, the decrease in the MLE variances improves faster than does the Bayesian. Compared to design 13, in design 17 the techniques had the same trends. The bias, variance and MSE for the location parameter for MLE and Bayesian went down. For b_3 they were all higher, as can be seen in Tables 3.7, 3.8, and 3.6, respectively. On the other hand, for b_2, the bias went up but the variance went down, lowering the MSE. The MLE has the lower bias for all parameters compared to Bayesian, when variances and MSEs are identical. In the

small sample case, the bias, variances, MSE of all parameters were lowest in the informative prior case.

Both regressors of design 21 are generated by c_4^2. This case produces results very similar to the above. All the bias, variance and MSE figures decrease compared to the small sample version, except for the bias of the informative Bayesian estimator for b_2. This result holds despite the fact that the X in Table 3.4 actually turns out to have less skew and lower kurtosis than they do in the small sample case in Table 3.4. Tables 3.3, 3.4, 3.5 and 3.6-3.8 present the risk figures from the experiment. Bias of the location parameter goes down in all instances, and it increases for b_3 compared to the normal-regressor case (Table 3.7). The bias of MLE estimate for b_2 goes up, whereas it goes down for Bayesian. Pattern for the variance is exactly same as the result patterns of design 17.

The large sample results also indicate that MLE improves, but regressor distribution has mixed effects. Recall that in design 9, where both regressors are skewed but the sample size is small, the informative Bayesian outperformed MLE in all measures of risk. Increasing sample size decreases risk for both the Bayesian and MLE, but the decrease in MLE is much larger.

To see the effects of an additional skewed regression in small and large samples, we can compare cases 17 and 21. When both regressors are skewed, the bias goes down even further for all parameters and for both Bayesian and MLE. Overall, the risk of design 21 is less than its counterparts in design 17. This result holds for large samples, too. However, as we have found earlier, as sample size increases, the performance of MLE improves faster.

3.8 Some Other Comparisons

One of the objectives of the Monte Carlo experiment is to see the effects of the sign of a parameter coefficient and correlation on the competing estimation processes. The former

comparison can be performed by looking at the design pairs 1-2, 5-6, 9-10, 13-14, 17-18, and 21-22 for low correlation data sets and the remaining consecutive pairs for high correlation. The description of data and the results being referred to in these comparisons are presented in Tables 3.3, 3.4, 3.5, 3.6, 3.7 and 3.8, and the sampling distributions are in Figures B.1-B.24.

Table 3.20: Descriptive Statistics for the Regressors Experimental Design 7 and 8

	X2	X3
Mean	0.1776	0.1915
Variance	1.0438	0.1489
Skewness	-0.3062	-0.3625
Kurtosis	-0.9233	-0.5082
Correlation	0.8152	0.8152
# of samples	500	500

Table 3.21: Descriptive Statistics for the Regressors Experimental Design 11 and 12

	X2	X3
Mean	0.0863	0.1185
Variance	0.9143	0.8969
Skewness	1.2697	0.9379
Kurtosis	0.2479	1.4247
Correlation	0.7651	0.7651
# of samples	500	500

Table 3.22: Descriptive Statistics for the Regressors Experimental Design 19 and 20

	X2	X3
Mean	-0.0148	0.0388
Variance	0.8062	1.0080
Skewness	0.0919	0.3430
Kurtosis	1.2239	0.1703
Correlation	0.7951	0.7951
# of samples	500	500

Table 3.23: Descriptive Statistics for the Regressors Experimental Design 23 and 24

	X2	X3
Mean	-0.0200	0.0346
Variance	0.8921	0.8796
Skewness	1.4204	0.8846
Kurtosis	2.8502	1.5844
Correlation	0.7574	0.7574
# of samples	500	500

Following the above pairwise comparisons, to look at the effects of the sign of the coefficient of b_3, we first observe the MSE factor of designs 1 and 2. The negative coefficient on b_3 increases the MSE for estimating the intercept but decrease it for the other parameters. These results can be seen in Tables 3.3 and 3.6. Bayesian estimator with lower prior variance outperforms the MLE in both designs. As expected, the biggest impact of changing the true parameter value on b_3 is on the risk measures of b_3. The results of the comparisons of designs 5-6 and 9-10 are very similar. MLE is more severely impacted when the coefficient value is positive. All the risk factors, bias, variance and MSE improve for b_2 when we have a positive true value for b_3, however, they all suffer in case of b_1 and b_3 as can be compared in Tables 3.6 and 3.3, respectively. The bad performance of MLE may be attributed to the fact that the sample size is small and 5, 6, 9, and 10 have at least one skewed X in the design. Observation of the remaining designs confirms the similar pattern. In case of design 13-14, the intercept and b_3 have increased estimation risk due to a positive true value (Table 3.6). However, in this case the impact is not as severe, due to a larger sample size. Bayesian and MLE results are similar, confirming the improvement speed of MLE following increases in the sample size. As the sign of b_3 changes from design 17 to 18, all the biases increase, however, variance and the MSE has the same pattern. In this case again, the impact on b_3 is severe. This confirms the previous

finding that when a regressor skewness and sign of the regressor coefficient are related. If a regressor is skewed to the right, i.e. the frequency mass toward larger positive values and the sign of the true parameter value is negative, the risk measures are impacted severely. On the other hand, given the right skewness, if the true parameter value is positive, this affects the estimation process favorably. We can see the same pattern comparing designs 21 and 22. We observe that Bayesian handles the change of the true value sign on a parameter better than Bayesian does. The same comparison when the correlation of Xs are high follows a similar trend but yields more mixed results as a result of competing effects. When the correlation is low, we have seen that when sample size is large, MLE has a smaller increase or larger decrease compared to its Bayesian counterparts. However, when the correlation is introduced the positive effects of sample size on the MLE are delayed. We observe the informative Bayesian outperforming MLE for both positive and negative b_3 when sample size is large. This was not the case when correlation was small. Also, the MSE of b_3 seem to increase when the sign is positive regardless of the estimation technique as expected.

For the effects of correlation on the performance of the estimation techniques, we can compare design pairs of 1-3, 5-7, 9-11, 13-15, 17-19, and 21-23 for negative coefficient of b_3 and 2-4, 6-8, 10-12, 14-16,18-20, and 22-24 for positive coefficient of b_3. The outcome of the observation is very conclusive. Introduction of correlation increases the MSE for almost all pairs, all parameters and all techniques, except for the location parameter for some instances. The overall impact on the location parameter is negligible (Table 3.6), whereas the impact on b_2 and b_3 (Tables 3.3 and 3.11) is considerable, usually doubling the MSE. Comparison of design 1 and 3 illustrates the point that the MSE of location parameter (Table 3.6) actually improves with

correlation whereas all the risk measures go up for b_3. The only difference when we compare designs 2-4 is that all the bias measures go down, however, not enough to outweigh the increase in the variances. A more interesting comparison is the comparisons of experimental designs 5-7- and 6-8. Although the pattern is similar, the magnitudes are very different. All the above mentioned designs incorporate one skewed regressor and the sample sizes are small. If we introduce correlation to design 5, all the MSE figures go up as well as the bias with the exception of informative Bayesian estimate for the location parameter as is tabulated in Table 3.6. The introduction of the correlation does not worsen one estimation process more than the other. On the other hand, when we incorporate correlation to design 6, MLE suffer much more than Bayesian estimation does. The resulting MSE of MLE is around three times that of the informative Bayesian MSEs.

Even in the large samples, where MLE does better, the introduction of the correlation increases the risk of the MLE to the extend that informative-Bayesian outperforms MLE.

One example of that is designs 21 vs. 23. In design 21, the MLE does at good as Bayesian for location parameter and better in for the other parameters than even informative Bayesian, however, as a result of introducing correlation, MLE suffers so much that informative Bayesian risk ends up less than that of the MLE. Similar result can be observed in design 22-24 comparison.

3.9 Conclusion

In this chapter a Monte Carlo experiment is designed for MLE and Bayesian probit models. The objective of the experiment is to investigate the sensitivity of Bayesian estimator to different non-informative priors, and also to asses the importance of changes in the factors such as true parameter value, regressor distribution, collinearity between the regressors, and sample

size on parameter estimates of the unconstrained probit model. The estimation techniques are compared based on the bias, variance, and MSE of the coefficient estimates.

The results of the experiments indicate that changing the prior variance improves the Bayesian estimation dramatically. This is despite the fact that there is not a big difference in the distribution. The MLE outperformed the Bayesian estimator with the larger variance in all experimental design points. The Bayesian estimator with the smaller prior variance, on the other hand, outperformed the MLE in almost all design points. The MLE improves faster when the sample size increase than does either of the Bayesian estimators. Collinearity adversely affect the estimates as expected. Bayesian technique seems to handle the collinearity better than MLE under certain conditions. Skewness of the distributions brings more risk to the estimation process. The risk increases if the regressors are from distributions with different degrees of skewness.

There are a large number of issues to be further investigated. The first issue is around the prior distribution. Since there is a big improvement in the Bayesian estimation as the prior variance changes, different non-informative and informative priors should be tried to look into the sensitivity of the risk factors to those changes. The experiment can also be improved by increasing the levels for each factor involved. In this Monte Carlo experiment there are two levels for each factor, which makes it hard to read the direction of the effects. A second round experiment can be run with 4-5 levels for factors that are found important in the first round. The results to the second round experiment could yield better read as to the direction and the significance of these effects. In addition to the existing factors, the experiment can also be extended by adding more factors such as stochastic regressors, low variability in the regressors

etc. In addition, the effects of these factors on the marginal probabilities as well as the elasticities and predictive probabilities can be observed.

CHAPTER 4

BAYESIAN POISSON MODELING

4.1 Introduction and Literature Review

The class of generalized linear models unifies the approaches needed to analyze data for which either the assumption of linear relation *x* and *y* or the assumption of normal variation is not appropriate. A generalized linear model is specified in three stages:

(i) The linear predictor, $h = Xb$,

(ii) The link function *g(.)* that relates the linear predictor to the mean of the outcome variable: $m = g^{-1}(h) = g^{-1}(Xb)$,

(iii) The random component specifying the distribution of the outcome variable *y* with mean $E(y \mid X) = m$. In general, the distribution of *y* given *x* can also depend on a dispersion parameter, f .

X is the *n x p* matrix of explanatory variables and $h = Xb$ is the vector of *n* linear predictor values. The sampling distribution takes the form;

$$p(y \mid X, b) = \prod_{i=1}^{n} p\left[y_i \mid (Xb)_i, f\right]$$

Typically, for the Poisson distribution, the dispersion parameter is fixed at 1. However, in many cases, there is significant overdispersion. The Poisson generalized linear model, also called the Poisson regression model, is used for count data problem. This model assumes that *y* is Poisson with mean and variance, m . The link function is taken to be the logarithm, indicating $\log m = Xb$. It is convenient to specify m as a log-

linear function of the explanatory variables that account for observed sample heterogeneity. In this setting, the systematic effects interact in a multiplicative way, and the coefficients have the interpretation of a partial elasticity of $E(y \mid x)$ with respect to the level of x if the logarithm of x is included among the regressors. The sampling distribution for data $y = (y_1, \ldots, y_n)$ becomes

$$p(y \mid b) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{-\exp(h_i)} \left[e^{h_i} \right]^{y_i}.$$

Due to the logartimic formulation, it can be shown that, $E\left[y \middle| x_i \right] = \exp(b' x_i)$.

To calculate the marginal effects, $\frac{\partial E\left[y \middle| x_i \right]}{\partial x} = \exp(b' x_i) b$. This will allow us to get directional reads about the effects of the coefficients.

The fact that the variance of the Poisson distribution is not independent of the mean poses questions regarding the flexibility of the Poisson regression model. However, it has been shown that the estimator of b remains consistent even if the variance does not equal the mean, indicating a distribution other than Poisson, as long as the link function is correctly specified. This robustness could be seen as analogous to a property of the linear model where OLS is unbiased independently of the second-order moments of the error distribution.

The notion of overdispersion results in the data exhibiting more variation than expected under the Poisson distribution due to systematic differences among subjects of interest in the study. Such variation could be incorporated in a hierarchical model using an indicator for each subject, with these indicators following a distribution. Hierarchical generalized linear models are a natural way to fit complex data structures and allow us to

include more explanatory variables without encountering the problems of overfitting. This generally provides larger standard errors of the estimated coefficients, as is required to correctly reflect the separate levels of variation.

Viallefont, Richardson and Green (2001) model the overdispersion using Poisson mixtures, with a variable "number of components" *k*, formalized as $y_i \sim \sum_{j=1}^{k} w_j f\left(. \mid m_j\right)$, where *w* denotes the weights and m the Poisson parameters in the mixture. The setup of the model can be viewed as a fully Bayesian method for model choice.

The initial Bayesian treatment of the Poisson regression is due to El-Sayyad (1973), who attempted to test the existence of a trend in the means of Poisson distributions. To this end, he introduces a Bayesian approximation for the posterior of the trend coefficient, and shows by numerical examples that the approximation works very well. He further postulates that even in Poisson experiments with small sample sizes the approximation could provide the initial solution to be used in the start of the maximum likelihood procedure.

Winkelmann (2000) presents a standard result of a closed form posterior distribution without covariates. Suppose $\{y_i\}, i = 1,\ldots,n$ is a random sample from a Poisson distribution with mean m, and the prior distribution of m is a gamma distribution with parameters $a \geq 0$ and $j \geq 0$. The gamma distribution is the conjugate prior for the Poisson parameter, and

$$g(m \mid y) \propto \left(\prod_{i=1}^{n} e^{-m} m^{y_i}\right) \frac{a^j}{\Gamma(a)} m^{a-1} e^{-mj}$$
$$\propto e^{-m(j-n)} m^{a+n\bar{y}-1}$$

Thus, the posterior distribution of m is a gamma distribution with parameters $\tilde{a} = a + n\bar{y}$ and $\tilde{j} = j + n$. The Poisson-gamma model is an example of a famous result in Bayesian analysis, namely that the posterior mean is a weighted average of prior mean and sample mean.

Chib and Winkelmann (2001) propose a class of models, whereby the correlation among the counts is represented by correlated latent effects to allow a general correlation structure. The types of correlated counts are defined to be;

(i) Genuine multivariate data on several related counted outcomes

(ii) Longitudinal measurements on a large number of subjects over a short period of time

(iii) Measurements on a small set of subjects over a long period of time, sometimes referred to as seemingly unrelated case

In their model, the counts are assumed to be independent Poisson with a conditional mean function that depends on the latent effects and a set of covariates. They further assume a multivariate Gaussian distribution for the latent effects with a zero mean vector and full unrestricted covariance matrix as well as multivariate-*t* distribution. They apply a certain kind of MCMC methodology due to Chib *et al.* (1998) to simulate the augmented posterior distribution of the parameters and the latent effect without computing the likelihood function of the model. The practicality of the approach in the case of higher dimensional problems is shown by examples.

In the seemingly unrelated case, King (1989) demonstrates a key difference between Poisson and linear regression models in that even when identical exogenous variables are used in both equations, a contemporaneous correlation among the

disturbances will generally yield a more efficient solution than equation-by-equation Poisson models. To the contrary, Zellner (1962) had shown just the opposite for the linear regression model, which reduces to the single equation LS estimators in the event of identical regressors even if the disturbance terms in different equations are correlated. In deriving his conclusions, King uses the property of the bivariate Poisson distribution that the zero covariance implies independence. He notes that the joint Poisson regression estimator provides a full information ML solution that is consistent and asymptotically more efficient than an equation-by-equation exponential Poisson model.

The issue of overdispersion is dealt with in a MCMC study, where Clyde and DeSimone-Sasinowska (1998) introduce a new approach for implementing Bayesian model averaging and sampling from large model spaces in the context of Poisson regression models using orthogonal or nearly orthogonal variables. The authors rely on a previously built argument that orthogonalizing the regressors can strongly improve convergence and mixing. The trade-off is to keep as many confounding variables as possible but at the same time come up with efficient schemes for finding models in large dimensional problems.

Given the empirical fact that most count datasets are overdispersed; i.e., sample variance is considerably larger than the sample mean, the generalized Poisson regression model seems to be the suitable choice as considered in Consul and Famoye (1992). One version of generalized Poisson distribution has the following probability function:

$$p(y \mid q, l) = q\,(q + y\,l)^{y-1}\,(y!)^{-1}\,e^{-q-y\,l} \qquad (4.1)$$

where $q > 0$ and $0 \le l < 1$. The mean and variance of this random variable are given by

$$E(y) = \frac{q}{1-l}$$

$$Var(y) = \frac{q}{(1-l)^3}$$

When $l = 0$, it reduces to the standard Poisson density with $E(y) = Var(y) = q$.

Another parametric model that deals with the overdispersion is negative binomial. Since the variance of negative binomial generally exceeds its mean, negative binomial deals with overdispersion better than the Poisson (Winkelmann, 2000).

In the case of the random count data believed to be affected by a number of explanatory variables, one adheres to a generalized Poisson regression model based upon the generalized Poisson density as defined in (4.1). Given the covariate vector x_i and the distribution of y_i be that of the GPD with mean $E(y_i \mid x_i; b, l) = m(x_i; b) = m_i > 0$. Furthermore, $m_i = \frac{q}{1-l} = qf$. Thus, the corresponding GPR model for the response variable y_i could be written as

$$P(Y_i = y_i \mid x_i; b, l) = m_i \left[m_i + (f-1) y_i\right]^{y_i - 1} f^{-y_i} (y_i!)^{-1} \exp\left\{-\left[m_i + (f-1) y_i\right] \div f\right\}$$

The parameter $f = \frac{1}{1-l}$ represents the square root of the index of dispersion, and that the variance of the response variable is $Var(Y_i \mid x_i; b, l) = f^2 m_i$.

The corresponding likelihood function will be of the following form:

$$l(q, l \mid Y) \propto q^n \exp(-nq - Z l) \prod_{i=1}^{n} \frac{(q + y_i l)^{y_i - 1}}{y_i!} \qquad (4.2)$$

where $Z = y_1 + \ldots + y_n$.

Given the flexibility of the shape of the Gamma distribution and the fact that its support lies between 0 and infinity on the real line, q is assumed to follow a Gamma distribution with parameter values a and b. The Gamma distribution enables one to impose relatively diffuse priors on q. In addition, $l \sim UNI(0,1)$ and is assumed to be independent from q. Then, the joint distribution of (q, l) is

$$P(q, l) \propto \frac{b^a}{\Gamma(a)} q^{a-1} \exp(-bq) \tag{4.3}$$

Combining (4.2) and (4.3) yields the posterior distribution of the parameters of the generalized Poisson regression model:

$$P(q, l \mid Y) \propto q^{n+a-1} \exp\left[-(n+b)q - Z l\right] \prod_{i=1}^{n} (q + y_i l)^{y_i - 1}$$

To make the application of the Gibbs sampler feasible, one could transform the posterior distribution by utilizing $l = qb$:

$$P(q, l \mid Y) \propto q^{Z+a} \exp\left[-(n+b)q - Zqb\right] \prod_{i=1}^{n} (1 + y_i b)^{y_i - 1}$$

where $q > 0$ and $0 \le b < q^{-1}$.

Thus, the full conditional distribution for q is given by;

$$P(q \mid Y; b) \propto q^{Z+a} \exp\left[-(n+b+Zb)q\right]$$

for $0 \le q < b^{-1}$. In other words, $P(q \mid Y; b) \sim Gamma(Z+a+1, n+b+Zb)$.

On the other hand, $P(b \mid Y; q) \sim$ Exponential(Zq) truncated on the interval $(0, q^{-1})$. The random draw for b could be obtained by the use of the inversion method

for truncated distributions. That is, one would sample from a *UNI(0,1)*, set

$U = \frac{1-\exp(-b\,Zq)}{1-\exp(-Z)}$, and solve for b.

4.2 Model

In this chapter we will not worry about the overdispersion and concentrate on the model that has been proposed by Winkelmann (2000). As mentioned earlier, a closed form posterior distribution exists for Poisson distribution. The Poisson-Gamma model represents the typical result in Bayesian statistics where the posterior is the weighted average of the prior and sample means. However, when the design matrix is introduced, the likelihood function becomes

$$L(b\,|y,x) \propto \prod_{i=1}^{n} \exp\left[-\exp(x_i'b)\right]\left[\exp(x_i'b)\right]^{y_i} \tag{4.4}$$

This expression is not a kernel of any distribution with either an informative or non-informative prior. There are a few proposed solutions, Albert and Pepple (1989) propose the use of approximation method. The other solution is employing methods that will simulate the exact distribution. In this chapter we will be simulating the exact posterior density via the Metropolis-Hastings (MH) algorithm (Chib and Greenberg, 1995).

Recall that the posterior density is:

$$p(b\,|y) = \frac{p(b)\;L(b\,|y,x)}{f(y)} \tag{4.5}$$

where $f(y) = \int_{\Theta} f(y|b)p(b)\,db$. Since that expression is not a function of b, we can express the posterior distribution as,

$$p(b|y) \propto p(b)\, L(b|y,x) \tag{4.6}$$

If we assume a constant as the diffuse prior, where $p(b) \propto c$, then the posterior will be proportional to the likelihood function in (4.4)

As mentioned earlier, since this is not the kernel of any known distribution, we will try to generate a sample from that exact distribution. Our target density (4.4) that we want to simulate can be obtained using the MH algorithm. The steps of the MH algorithm are;

(1) Pick the proposed density and evaluate it at the starting values for the unknown parameters.

(2) Draw a candidate b^* from $N(b_{(m)}, cV)$ where c is the scalar to control the accept/reject ratio r where $r = \frac{f(b^*|y)}{f(b_{(m)}|y)}$.

(3) Check if the candidate is satisfying the inequality restriction, if not resample.

(4) If r > 1, then the proposed is accepted since it improves the objective function and set $b_{(m+1)} = b^*$, where we assign the proposed value to a sample value.

(5) Draw a random uniform variable from U(0,1).

(6) If $u \leq r < 1$, then set $b_{(m+1)} = b^*$

(7) If u > r then the chain remains in its current position and samples again.

(8) The acceptance/rejection algorithm continues until convergence,

(9) A given number of samples are discarded to break the dependence to the starting values. The resulting sample is a sample from the exact target density.

The algorithm works in the following fashion: A candidate value is drawn from the candidate distribution. A rejection rate is calculated as the ratio of the where the distribution will be and where it is now. If the ratio is greater than 1, that means the algorithm is improving, thus the value is accepted and the algorithm moves to the new point. If the ratio is less than one, then the algorithm moves to that point with the probability equal to the calculated rate.

The most crucial part of the algorithm is the choice of the proposal density. Chib *et al.* (1998) suggest that, the proposed density for Poisson regression can be chosen as the normal distribution. Once the sample from the target density is obtained, it is straightforward to calculate any distributional characteristic or any function of these characteristics.

Our contribution to the literature by this chapter is to introduce inequality constraints on the parameters of the Poisson regression. This has been done by dropping any sampled value that fall in the restricted area before the calculation of the accept/reject ratio. The exact sample obtained provides all the information we need unlike the Maximum Likelihood Estimation (MLE) where it would be extremely tedious to evaluate the standard errors after such truncation.

4.3. Application

Cameron and Trivedi (1986), Cameron, Trivedi, Milne and Piggott (1988), Cameron and Trivedi (1993) used Poisson Regression model to estimate the Demand for Health Care and Health Insurance in Australia. We got the data from Cameron and Trivedi (1998). The sample of 5,190 of single individuals older than 18 has been derived from the original sample of 40,650 individuals taken from the A.B.S. 1977-78 Australian

Health Survey. The frequency distribution of the dependent variable, DVISITS is summarized in Table 4.1.

Table 4.1: Frequency Distribution of the number of consultations with a doctor or specialist in the past 2 weeks.

	0	1	2	3	4	5	6	7	8	9
# of Visits	4141	782	174	30	24	9	12	12	5	1
% of Overall Visits	79.79	15.07	3.35	0.58	0.46	0.17	0.23	0.23	0.10	0.02

The data set includes socioeconomic, status of health insurance of the individuals as well as their recent and long term health measures. We consider all these variables as did Cameron and Trivedi (1998). Definitions and descriptive statistics for the independent and dependent variables are presented in Table 4.2. As can be seen from the table, the independent variables are composed of binary, count, as well as continuos variables. We propose to apply the MH methodology to this model and introducing inequality restrictions on the coefficients of the socioeconomic variables. We suggest positive values for the coefficients of sex, and age and negative coefficients on income and age squared. These signs indicate that the females visit the doctors' office more and the number of visits increases with age and decrease by income. Given the Poisson Regression Model the vector of explanatory variables x_i is of dimension 13 and the components are as follows:

$x_{1=}$ 1;

$x_{2=}$ 1 if female, 0 if male (SEX);

$x_{3=}$ Age (AGE);

$x_{4=}$ Age Squared (AGESQ);

$x_{5=}$ Income (INCOME);

$x_{6=}$ If covered by private health insurance or not (LEVYPLUS);

$x_{7=}$ If covered by government or not (FREEPOOR);

$x_{8=}$ If covered free by government or not (FREEPERA);

$x_{9=}$ Number of illnesses in past 2 weeks (ILLNESS);

$x_{10=}$ Days with reduced activity due to illness or injury (ACTDAYS);

$x_{11=}$ Goldberg's health questionnaire score (HSCORE);

$x_{12=}$ If chronic condition(s) and not limited activity (CHCOND1);

$x_{13=}$ If chronic condition(s) and limited activity (CHCOND2);

The inequality restrictions are not only intuitive, but they are also supported by the earlier research. The signs imposed are consistent with the Poisson MLE estimates. The coefficients are not the marginal effects but they can give us the directional effect. As mentioned before, given an exponential link function, the effect of a unit change in any given regressor will effect the dependent variable by the product of the coefficient and the value of the link function evaluated at a given observation and that coefficient value and can be shown as $\frac{\partial E[y|x]}{\partial x} = b_j \exp(b'x_i)$. Since the exponential function is always positive, the sign of the coefficient indicates the direction of the effect.

Table 4.2: Descriptive Statistics of All Variables in the Dataset

Variable	Mean	SD	Skewness	Kurtosis	Minimum	Maximum
SEX	0.521	0.500	-0.100	1.000	0.000	1.000
AGE	0.406	0.205	0.400	1.500	0.190	0.720
AGESQ	0.207	0.186	0.600	1.800	0.036	0.518
INCOME	0.583	0.369	0.700	2.800	0.000	1.500
LEVYPLUS	0.443	0.497	0.200	1.100	0.000	1.000
FREEPOOR	0.043	0.202	4.500	21.400	0.000	1.000
FREEREPA	0.210	0.408	1.400	3.000	0.000	1.000
ILLNESS	1.432	1.384	0.900	3.200	0.000	5.000
ACTDAYS	0.862	2.888	3.800	16.800	0.000	14.000
HSCORE	1.218	2.124	2.400	9.300	0.000	12.000
CHCOND1	0.403	0.491	0.400	1.200	0.000	1.000
CHCOND2	0.117	0.321	2.400	6.700	0.000	1.000
DVISITS	0.302	0.798	4.700	34.300	0.000	9.000

4.4 Bayesian Estimation with Inequality Restrictions

Using the Australian Demand for Health Care and Health Insurance dataset, and Bayesian methodology, two models are estimated. The first model used MH algorithm to simulate the exact posterior distribution. In this model the parameter space has not been restricted. The same methodology has been used for the second model with an additional step included in the algorithm. This additional step checks if the proposal values for the parameters are within the restricted region of the parameter space or not. If the values sampled are outside of the restricted parameter space, they are excluded. This constrained version of the algorithm produces a sample that is restricted to the negative or positive side of the parameter space based on our specification. A diffuse prior that is proportional to a constant has been chosen for both of the models. The results of the estimation process for unconstrained Bayesian, MLE and constrained Bayesian Poisson

estimation are in Tables 4.3, 4.4 and 4.5, respectively. The tables include the means, and standard deviations.

Since Poisson model fundamentally is a nonlinear regression, we can estimate it using nonlinear least squares (NNLS) and use the NLLS estimates for b^* as the starting values for the MH algorithm that will facilitate the simulation.

The NLLS estimates produce good starting values in fewer iterations. In order to capture a 40% accept/reject rate we selected a scalar multiple of 0.45 by trial and error to multiply the ML covariance matrix. The objective of which is to shrink the covariance matrix in an effort to constrain the algorithm from roaming around. The algorithm produced 200,000 accepted samples and the first 40,000 have been discarded as part of the burn-in period.

Inequality restrictions are imposed on the coefficient of SEX, AGE, AGESQ, and INCOME variables.

Graphical convergence checks have been observed for all of the coefficients for both models. Some of these figures can be found in the appendix C. Figures C.1 - C.4 suggest that all the chains have achieved stationarity.

Figures 4.1 – 4.4 illustrates the posterior density for unconstrained and constrained models for coefficients of SEX, AGE, AGESQ and INCOME, respectively.

As can be seen from the Tables 4.3 and 4.4, the unconstrained Bayesian estimation yields results that are very similar to those of MLE. Since in the absence of prior information the posterior is proportional to the likelihood function, this result is very intuitive. On the other hand, comparison of Tables 4.3 and 4.5 indicate that the

coefficients on Age and Agesq had the largest impact. This is due to the high correlation between the two variables.

Table 4.3 : Bayesian Estimates for the Dr. Visits Data – Unconstrained Model

VARIABLE	Mean	Std Dev	Minimum	Maximum
CONSTANT	-2.2271	0.1894	-3.0784	-1.4168
SEX	0.1587	0.0563	-0.1082	0.4024
AGE	1.0462	1.0029	-3.0524	5.4469
AGESQ	-0.8386	1.0812	-5.5821	3.8294
INCOME	-0.2038	0.0889	-0.5736	0.1454
LEVYPLUS	0.1235	0.0723	-0.1888	0.4299
FREEPOOR	-0.4496	0.1801	-1.2785	0.2796
FREEREPA	0.0809	0.0920	-0.3135	0.4578
ILLNESS	0.1869	0.0182	0.1169	0.2596
ACTDAYS	0.1268	0.0050	0.1054	0.1506
HSCORE	0.0301	0.0100	-0.0103	0.0702
CHCOND1	0.1136	0.0670	-0.1474	0.3914
CHCOND2	0.1384	0.0832	-0.2019	0.4545

Table 4.4 : MLE Estimates for the Dr. Visits Data – Unconstrained Model

VARIABLE	Mean	Std Dev
CONSTANT	-2.2240	0.1900
SEX	0.1570	0.0560
AGE	1.0560	1.0010
AGESQ	-0.8490	1.0780
INCOME	-0.2050	0.0880
LEVYPLUS	0.1230	0.0720
FREEPOOR	-0.4400	0.1800
FREEREPA	0.0800	0.0920
ILLNESS	0.1870	0.0180
ACTDAYS	0.1270	0.0050
HSCORE	0.0300	0.0100
CHCOND1	0.1140	0.0660
CHCOND2	0.1410	0.0830

On the other hand, the coefficients of sex and income didn't change as much. By observing the extreme values of the constraint variables as well as their distributions, it can be seen that the constraints on the coefficients of AGE and AGESQ are the most binding ones. That also explains the huge reduction in the posterior standard deviation as

the constraint is imposed. The posterior standard deviations of the constrained model are also smaller than those of the MLE. As expected the restrictions increased the absolute value of the other coefficients and decrease their posterior standard deviation as well but the magnitude is negligible.

Table 4.5: Bayesian Estimates for the Dr. Visits Data – Constrained Model

VARIABLE	Mean	Std Dev	Minimum	Maximum
CONSTANT	-2.2848	0.1556	-3.0871	-1.7556
SEX	0.1582	0.0555	0.0000	0.4173
AGE	1.4091	0.7492	0.0004	4.9131
AGESQ	-1.2347	0.7986	-5.0266	-0.0002
INCOME	-0.2133	0.0855	-0.5733	0.0000
LEVYPLUS	0.1218	0.0720	-0.1570	0.3989
FREEPOOR	-0.4515	0.1779	-1.1695	0.2063
FREEREPA	0.0794	0.0921	-0.3034	0.4548
ILLNESS	0.1875	0.0184	0.1121	0.2710
ACTDAYS	0.1269	0.0050	0.1061	0.1489
HSCORE	0.0300	0.0101	-0.0081	0.0698
CHCOND1	0.1111	0.0673	-0.1466	0.3974
CHCOND2	0.1334	0.0830	-0.1933	0.4868

As mentioned earlier these coefficients are not the effects of a unit change in the dependent variables but rather signals of the direction of the effect. In order to calculate the marginal effects, we need to evaluate the effect at certain observations. We picked three observations, namely observation 766, 1244, and 2678. These observations are extreme observations where all the variables are at their minimum when observation 1244 is concerned. Observation 766 has maximum overall value except for the income. Observation 2678 has the maximum values for the constrained variables. The marginal effects for these observations for some of the coefficients are presented in Table 4.6 for both unconstrained and constrained models. The values for the marginal effects also indicate significant changes in the effects of the AGE and AGESQ variables. The

magnitude of the change is very similar to the magnitude of change in the coefficient estimates. The increase in the absolute value of the coefficients as a result of the restrictions is about 34% for the AGE variable and 47% for AGESQ. The change in the marginal effect of the AGE variable in different observations is 34%, 32%, and 34% for observations 1244, 2678, and 776, respectively. The change for the AGESQ is 47%, 43%, and 46% for the observations 1244, 2678, and 776, respectively.

Table 4.6 Marginal Effects of the Coefficients for some variables in both Models

MODEL	Unconstrained Model			Constrained Model		
OBSERVATION	766	1244	2678	766	1244	2678
AGE	3.89	0.13	0.25	5.22	0.18	0.33
AGESQ	-3.12	-0.11	-0.20	-4.57	-0.16	-0.29
INCOME	-0.76	-0.03	-0.05	-0.79	-0.03	-0.05

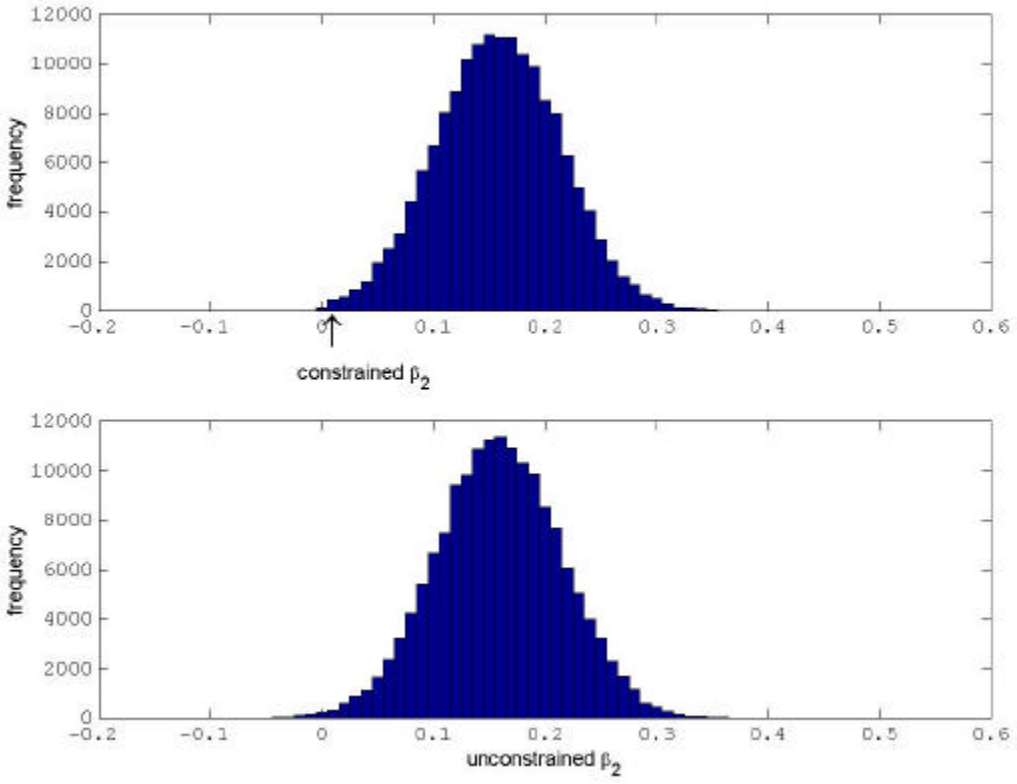


Figure 4.1: Posterior pdfs for Coefficient of the SEX variable in the unconstrained and Constrained Models

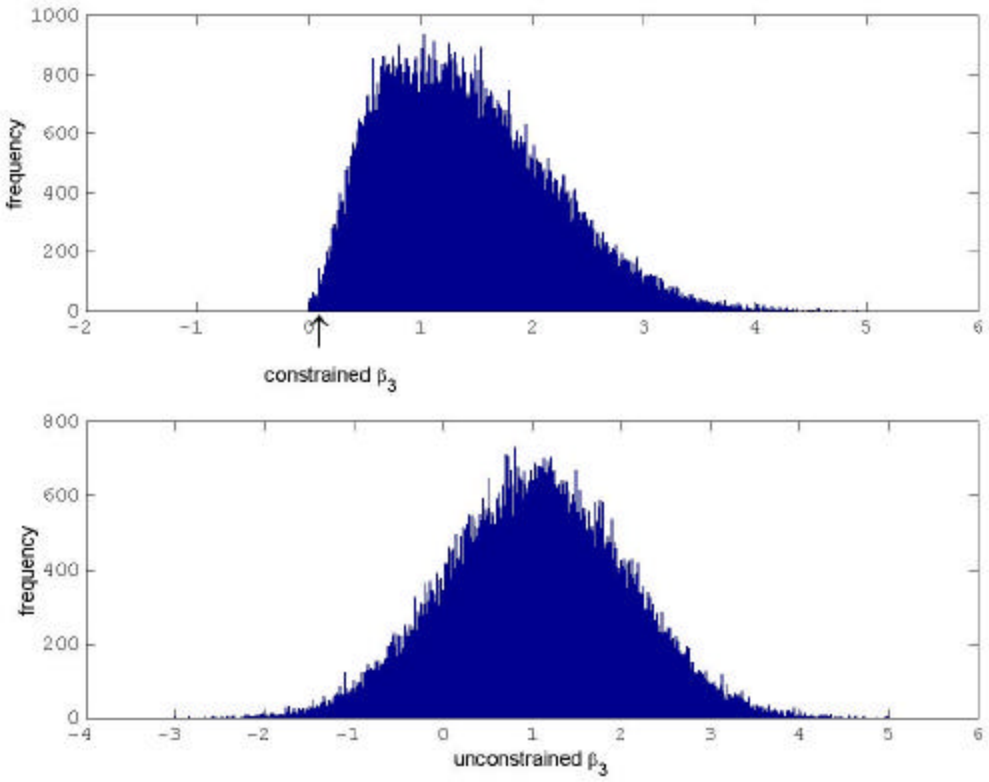


Figure 4.2: Posterior pdfs for Coefficient of the AGE variable in the unconstrained and Constrained Models

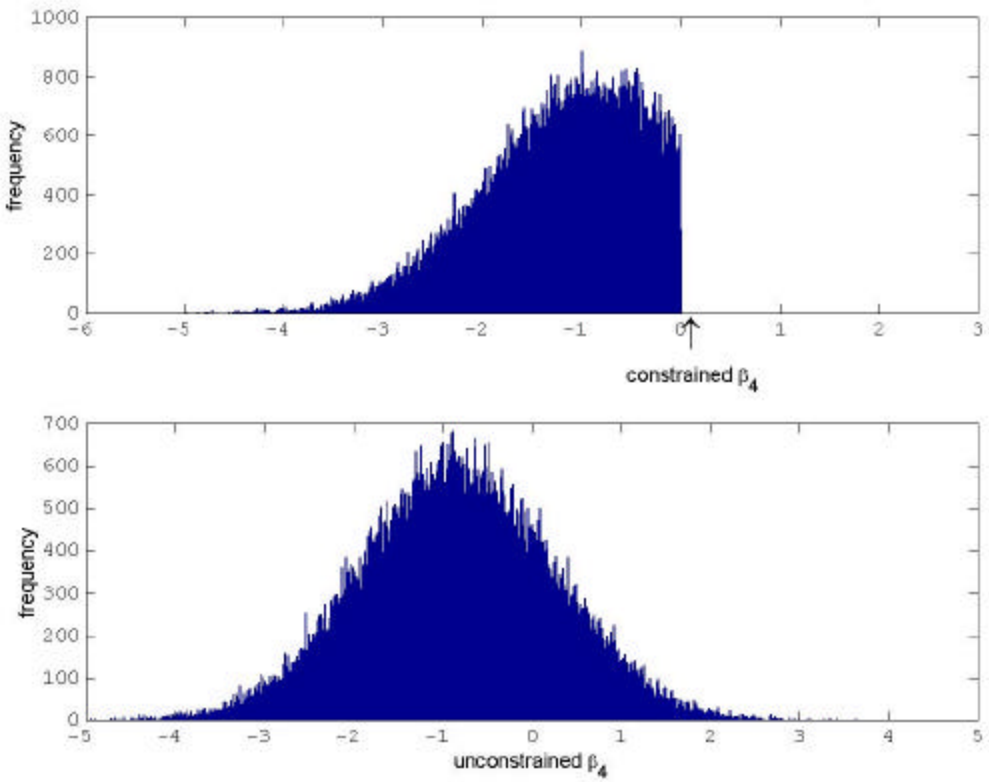


Figure 4.3: Posterior pdfs for Coefficient of the AGESQ variable in the unconstrained and Constrained Models

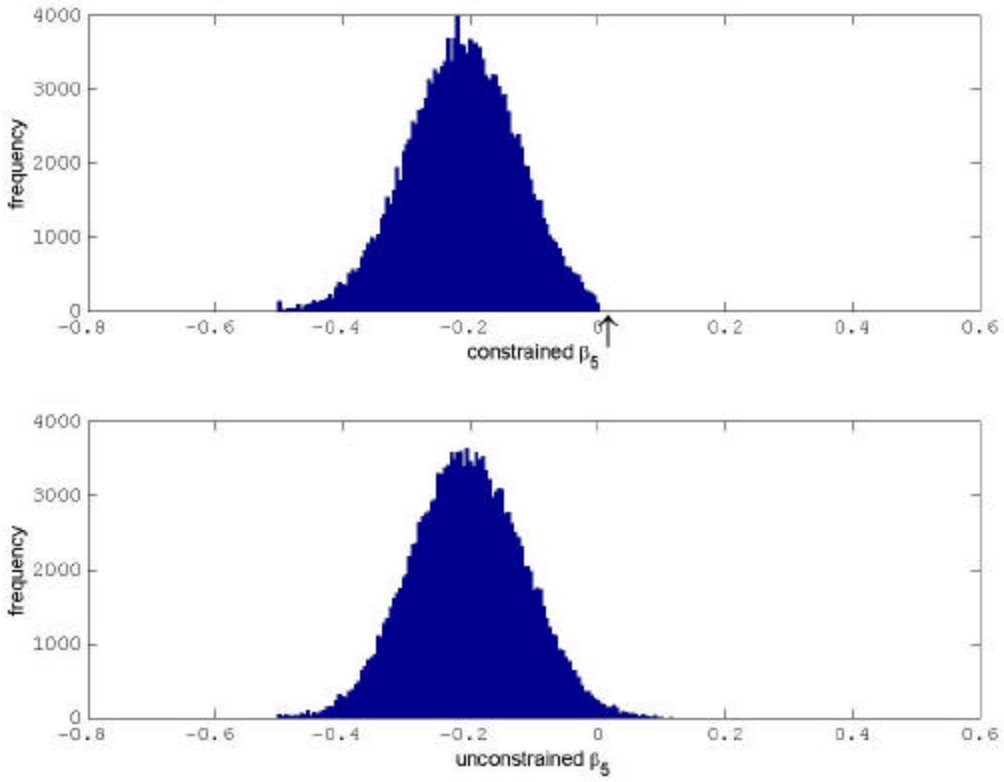


Figure 4.4: Posterior pdfs for Coefficient of the INCOME variable in the unconstrained and Constrained Models

4.5 Conclusion

Maximum Likelihood estimation of poisson model has been used in econometrics extensively. Bayesian applications of the poisson model has also been used frequently in the last few years. In this chapter, we estimated the poisson model with Bayesian technique using Metropolis algorithm. The contribution of this chapter is to introduce inequality constraints on the parameters of the poisson regression.

The exact sample obtained via the Bayesian estimation technique will provide us with all the information we need, unlike the Maximum Likelihood Estimation (MLE) where it would be extremely tedious to evaluate the standard errors after such truncation (Winkelmann, 2001).

The data used for this application is a sample of 5,190 of single individuals older than 18 that has been taken from the A.B.S. 1977-78 Australian Health Survey (Cameron and Trivedi, 1998). We are going to use this data to estimate the number of doctor visits in a period of 2 weeks. The independent variables in the survey can be combined into 3 categories as socioeconomic, insurance and health status variables. We propose to use truncated priors for the socioeconomic variables.

The ML estimate of this model is the value that maximizes the log likelihood function. Newton-Raphson algorithm is used to solve this nonlinear optimization. Since the conditional densities are not straightforward for Poisson regression, we utilized Metropolis Algorithm for Bayesian estimation. As expected, I observed an efficiency gain from incorporating nonsample information. The posterior standard deviations of the restricted model are smaller than those of unrestricted Bayesian and MLE.

One extension to this study can be to use negative Binomial distribution instead of the poisson distribution. Since the variance of the negative binomial is larger than its mean, it might be better able to take care of the overdispersion in the dataset. Another important issue is the effect of the restrictions on the marginal effects. Predictive power of the restricted Bayesian estimator can be observed and compared to that of the unrestricted Bayesian and MLE.

4.6 References

Albert J., Chib S. (1996) "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo," Edt. In *Advances in Econometrics* Volume 11 Part A, 3-24.

Arjas, E., Heikkinen J. (1997) "An algorithm for Nonparametric Bayesian Estimation of a Poisson Intensity," *Computational Statistics*, 12, 385-402.

Best, N. G., Spiegelhalter D. J., Thomas A., & Brayne, C. E. G (1996) "Bayesian

Analysis of Realistically Complex Models," *Journal of the Royal Statistical Society A*, 159, 323-342.

Brooks, S. P. (1998) "Markov Chain Monte Carlo Method and its Application," *The Statistician*, 47, 69-100.

Cameron, C., and Trivedi, P. K. (1998) *Regression Analysis of Count Data*, Cambridge, U.K: Cambridge University Press.

Cameron, C., and Trivedi, P. K. (1993) "Test of Independence in Parametric Models: with Applications and Illustrations," *Journal of Business and Economic Statistics*, 11, 29-43.

Cameron, C., and Trivedi, P. K. (1986) "Econometric Models Based on Count Data: Comparison and Application of Some Estimators," *Journal of Applied Econometrics*, 1, 29-53.

Cameron, C., Trivedi, P. K., Milne, F., and Piggott, J. (1988) "A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia," *Review of Economic Studies*, 55, 85-106.

Chib, S., and Winkelmann, Rainer (2001) "Markov Chain Monte Carlo Analysis of Correlated Count Data," *Journal of Business & Economic Statistics*, 19, 428-435.

Chib, S., Greenberg, E. and Winkelmann, R. (1998) "Posterior Simulation and Bayes Factors in Panel Count Data Models" *Journal of Econometrics*, 86, 33-54.

Chib, S., Greenberg, E. (1996) "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, 49, 327-335.

Chib, S., Greenberg, E. (1995) "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327-335.

Christiansen C. L., Morris, C. N. (1997) "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618-632.

Clyde, M., Desimone-Sasinowiska H. (1998) "Accounting for Model Uncertainty in Poisson Regression Models: Particulate Matter and Morality in Birmingham, Alabama," Institute of Statistics and Decision Science Discussion Paper 97-06.

Doss, H., Narasimham B. (1994) "Bayesian Poisson Regression using the Gibbs Sampler: Sensitivity Analysis through Dynamic Graphics," Working Paper.

El-Sayyad, G. M. (1973) "Bayesian and Classical Analysis of Poisson Regression," Working Paper.

Griffiths, E. William, Hill, R. Carter, and O'Donnell J. Christpher (2001) "Including Prior Information in Probit Model Estimation," Working Paper

King, G. (1989) "Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator," *American Journal of Political Science*, 33,762-784.

Oh, M., Lim, Y. B. (2001) "Bayesian Analysis of time Series Poisson Data," *Journal of Applied Statistics*, 28, 259-271.

Smith, A. F. M., Roberts G. O. (1993) "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society B*, 1, 3-23.

Viallefont, V., Richardson, S. and Green, P. J. (2000) Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics*, to appear.

Winkelmann, R. (2000) *Econometric Analysis of Count Data* (3rd ed.) Heidelberg: Springer-Verlag.

Zellner, A. (1962) "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 1962, 348-368.

APPENDIX A ADDITIONAL PLOTS FOR CHAPTER 2

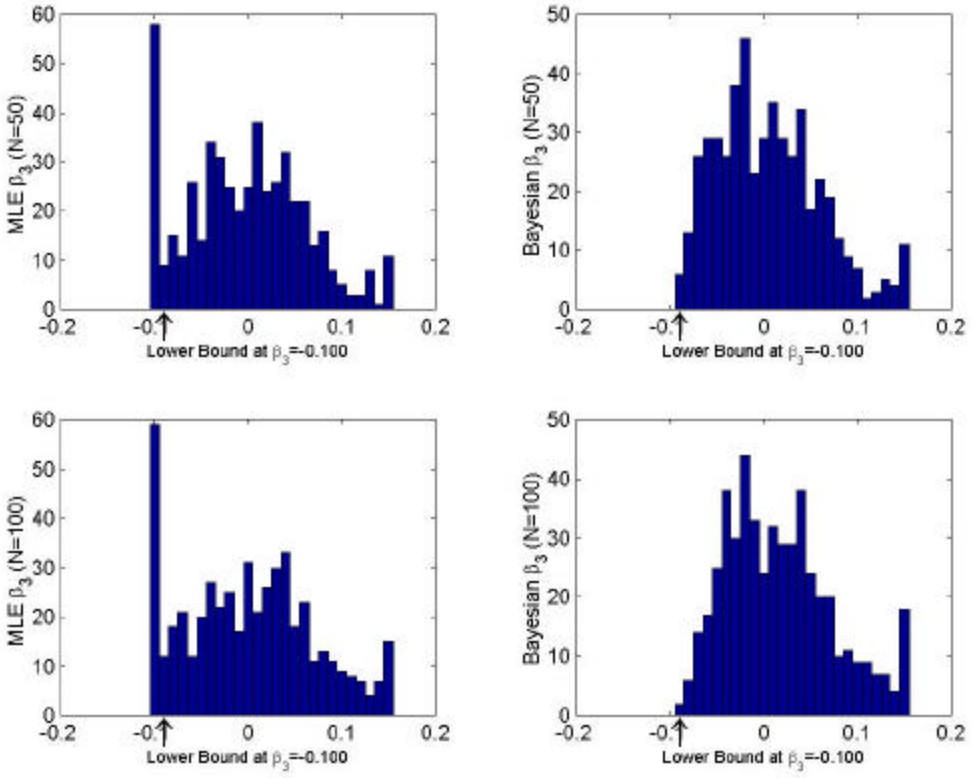


Figure A.1: Sampling Distributions with a lower bound of –0.100

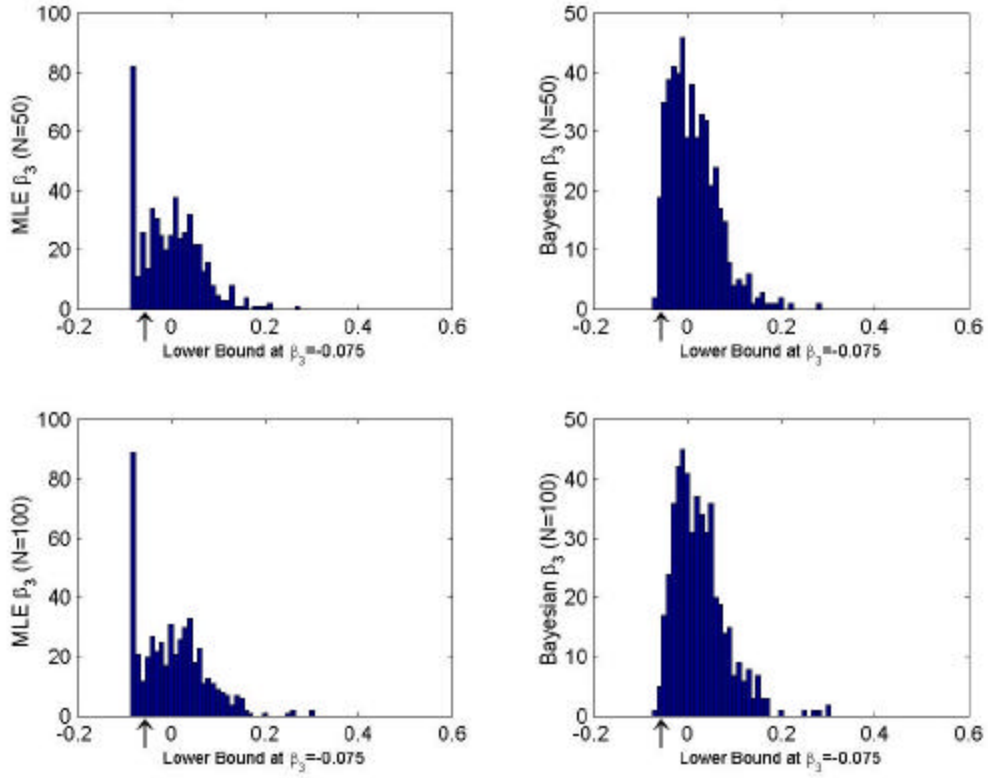


Figure A.2: Sampling Distributions with a lower bound of –0.075

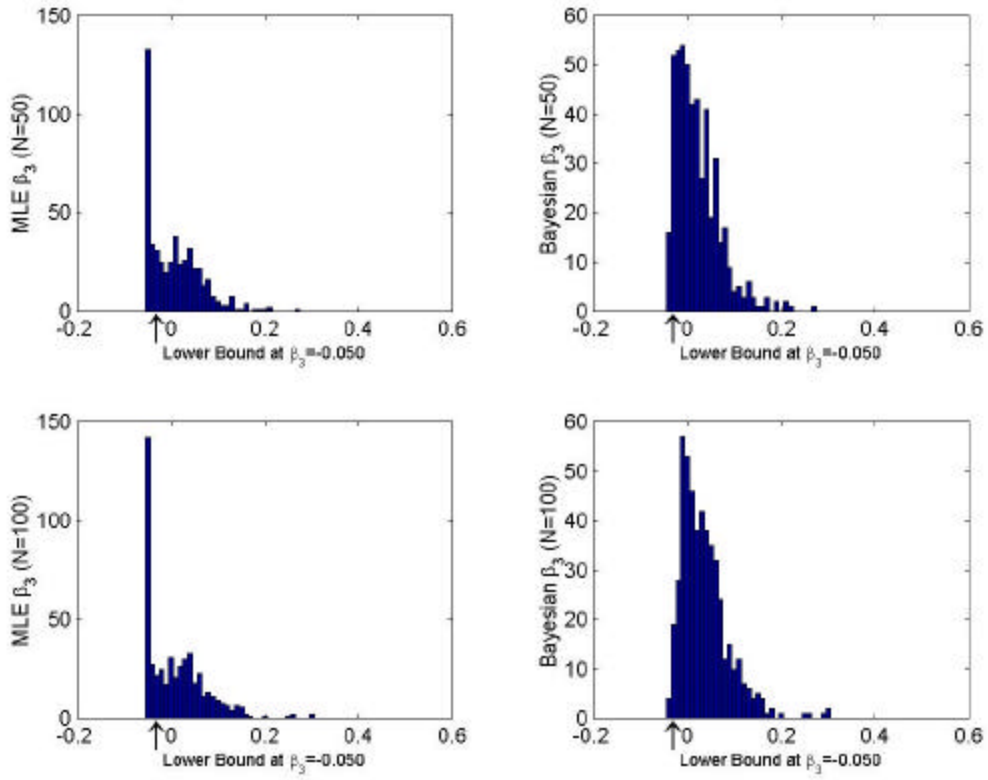


Figure A.3: Sampling Distributions with a lower bound of –0.050

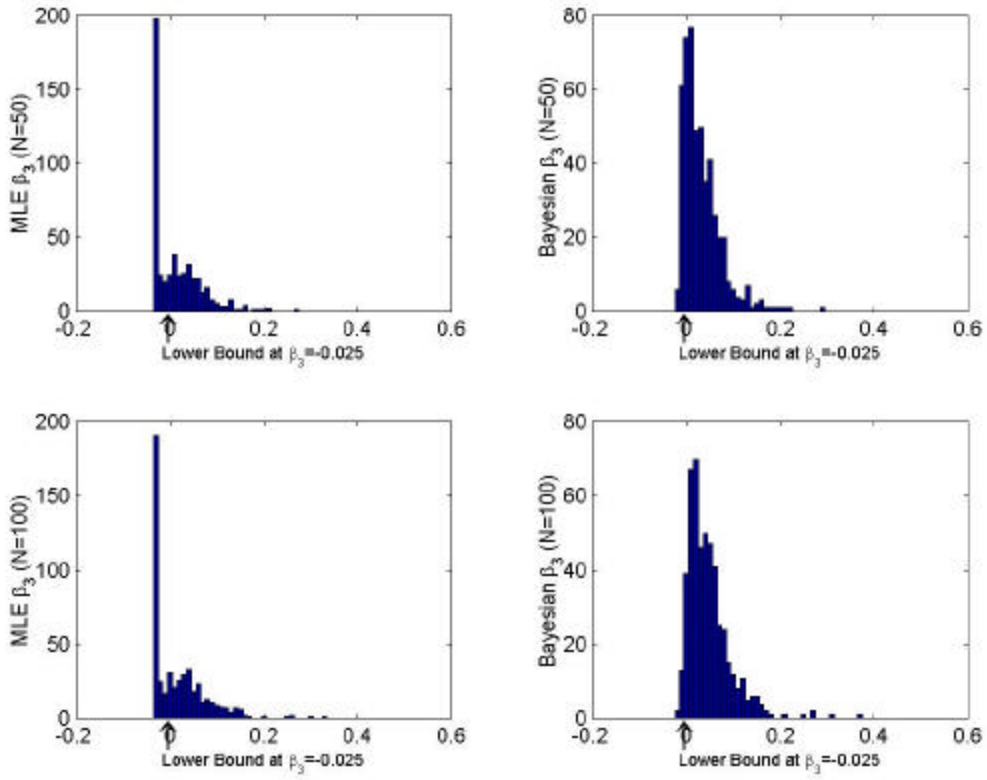


Figure A.4: Sampling Distributions with a lower bound of –0.025

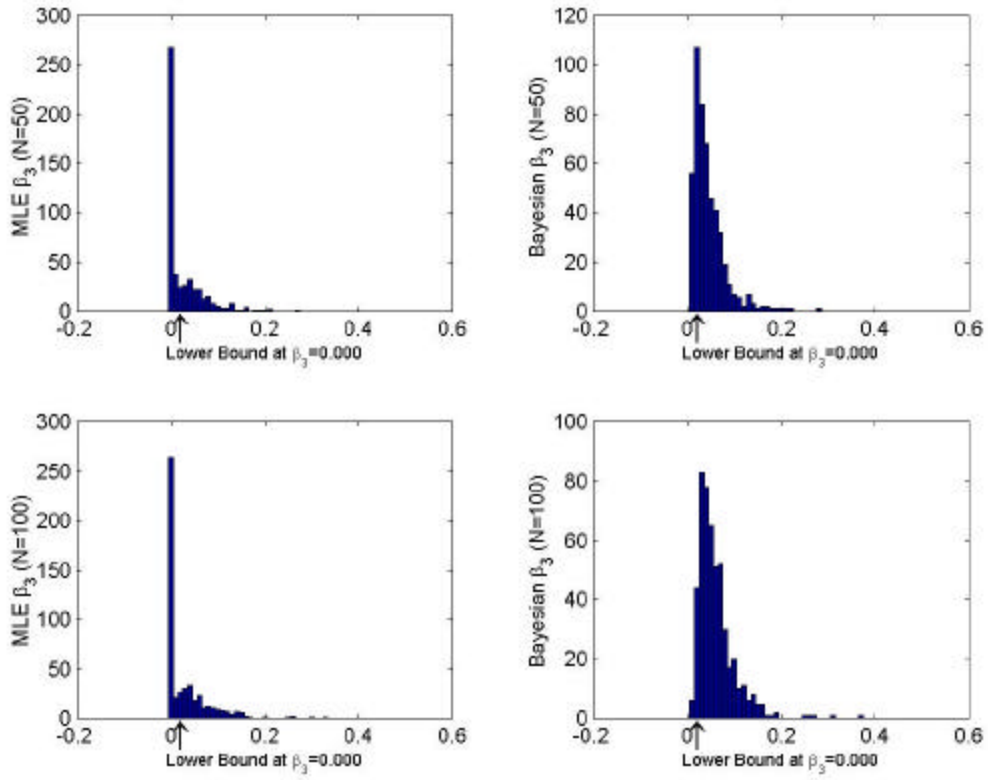


Figure A.5: Sampling Distributions with a lower bound of 0.000

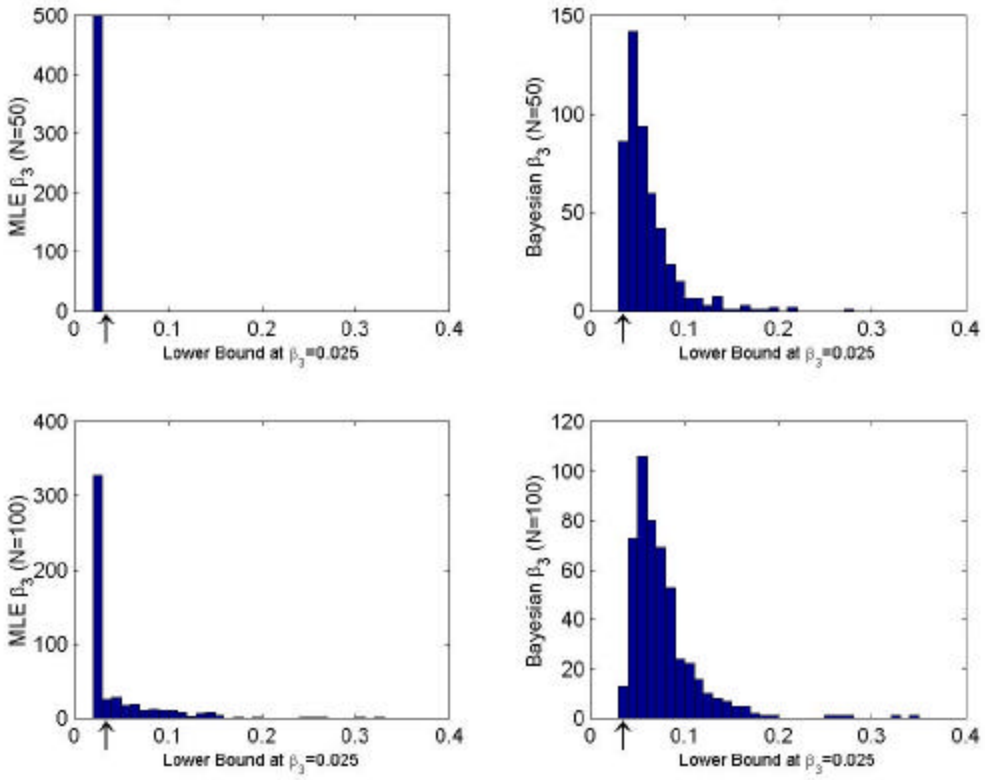


Figure A.6: Sampling Distributions with a lower bound of 0.025

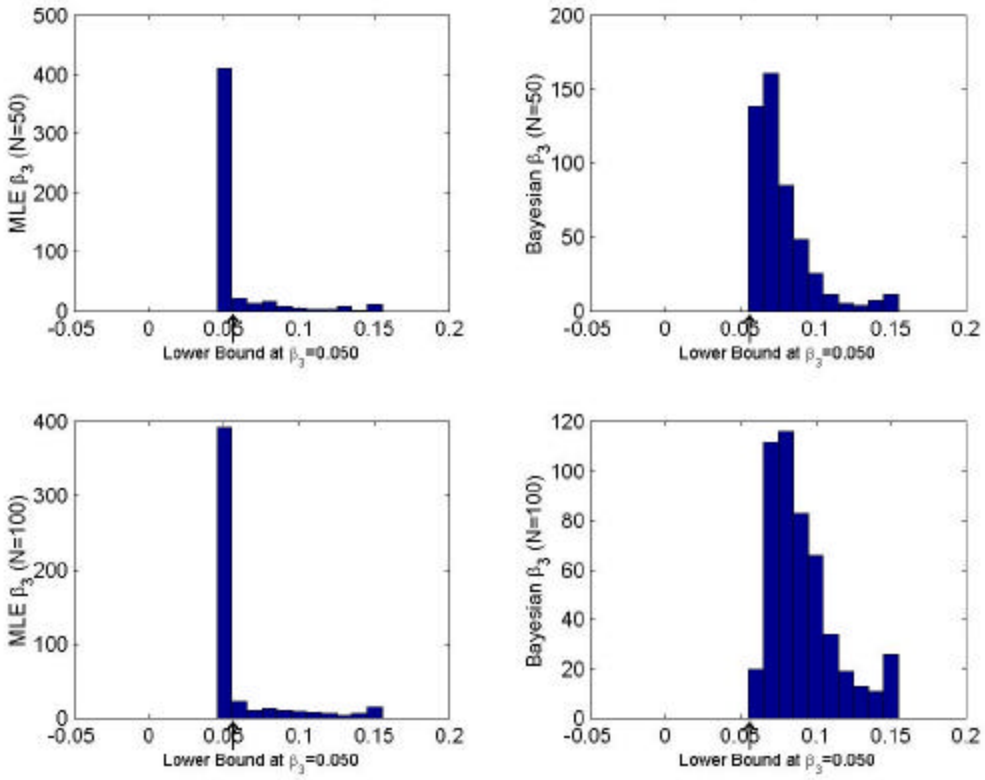


Figure A.7: Sampling Distributions with a lower bound of 0.050

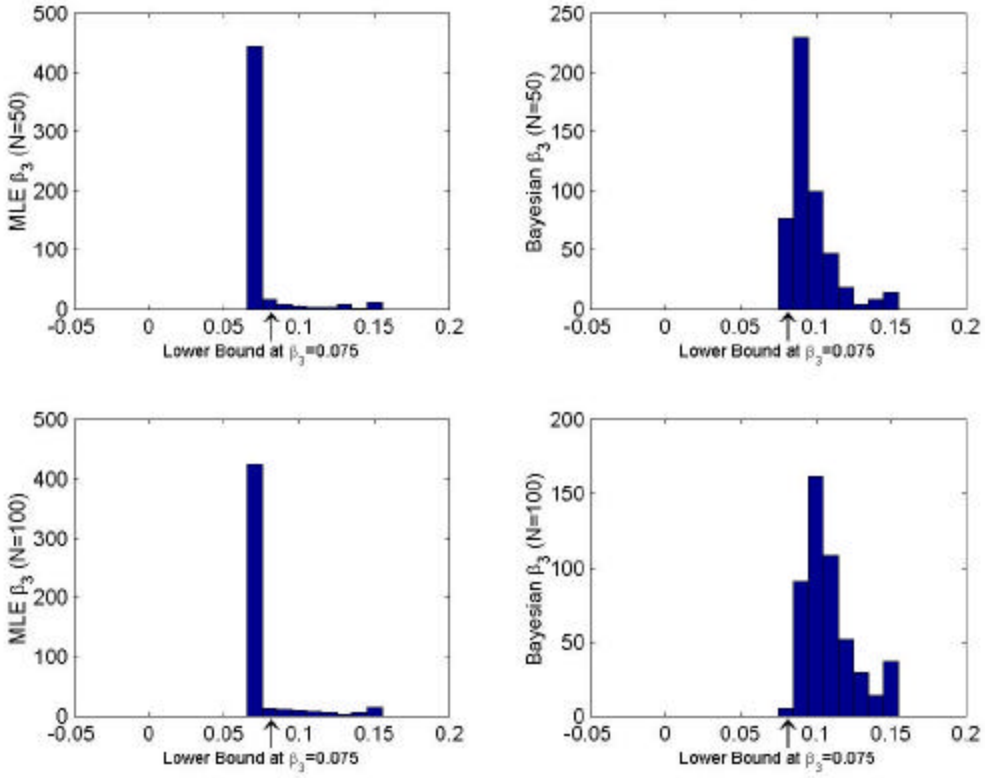


Figure A.8: Sampling Distributions with a lower bound of 0.075

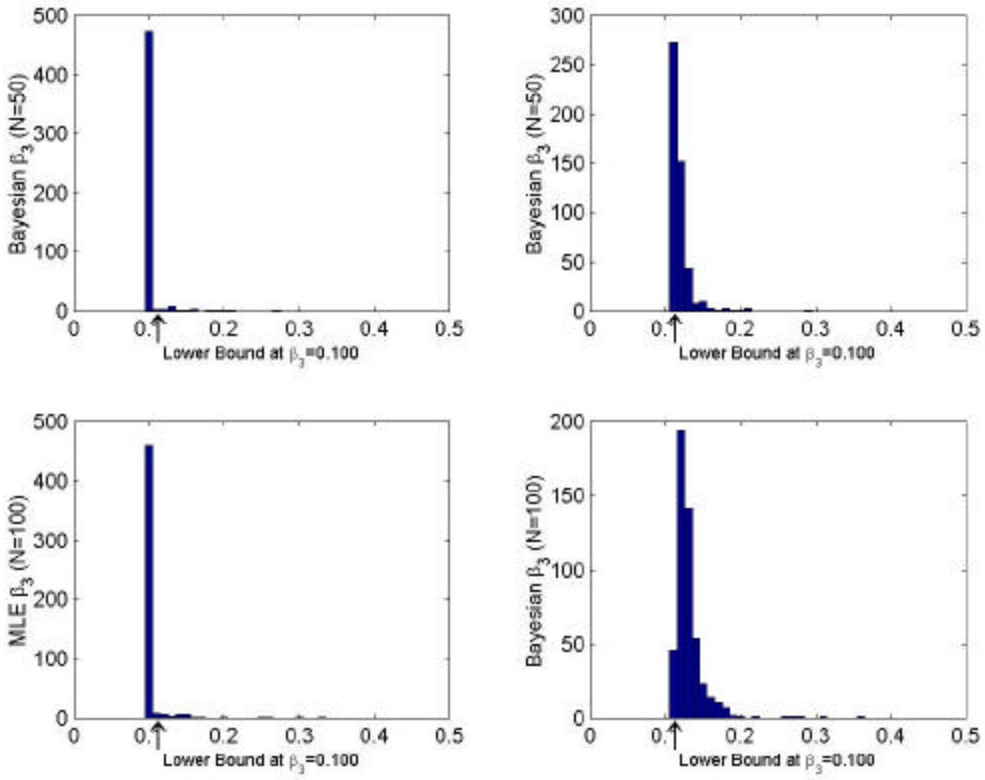


Figure A.9 Sampling Distributions with a lower bound of 0.100

APPENDIX B ADDITIONAL PLOTS FOR CHAPTER 3

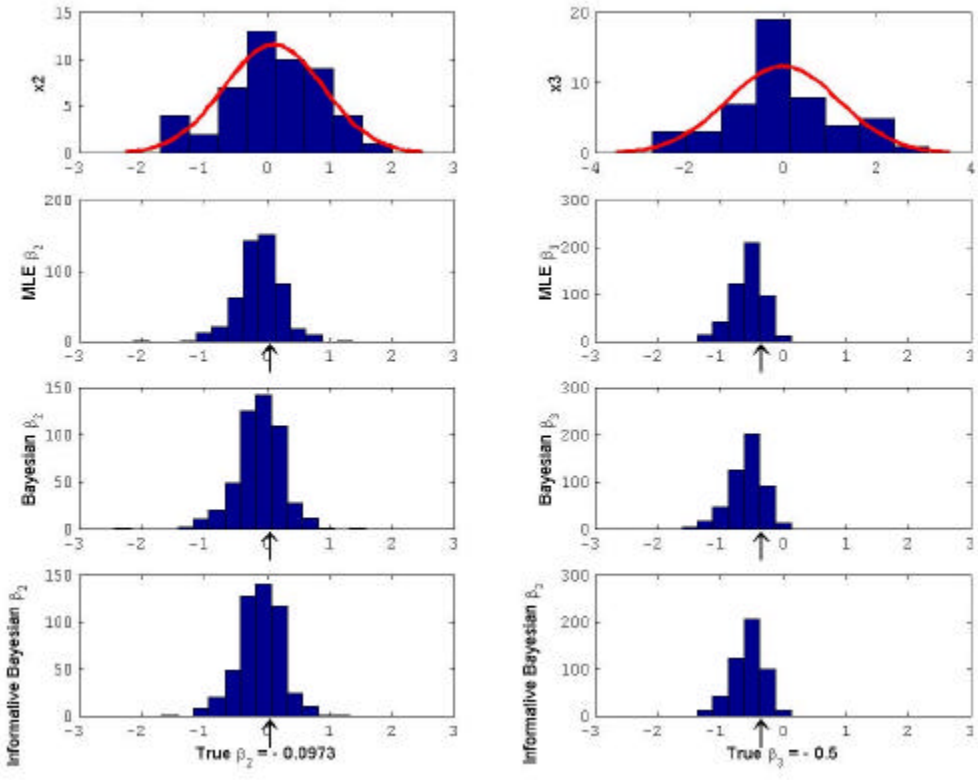


Figure B.1: Sampling Distributions for Design Point 1

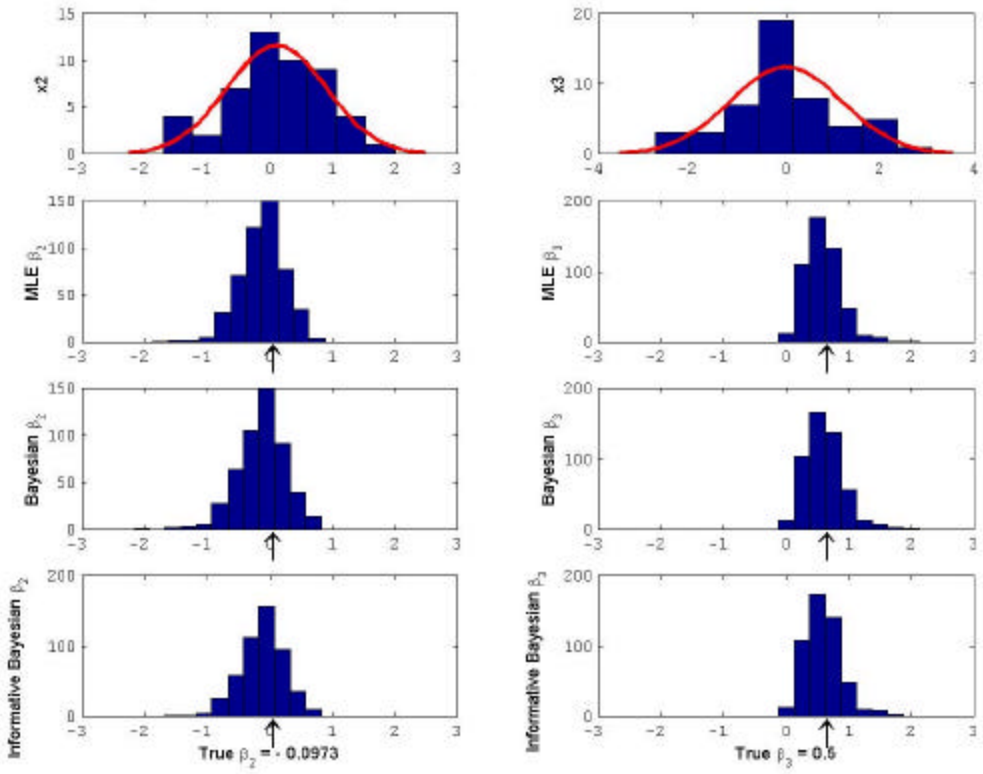


Figure B.2: Sampling Distributions for Design Point 2

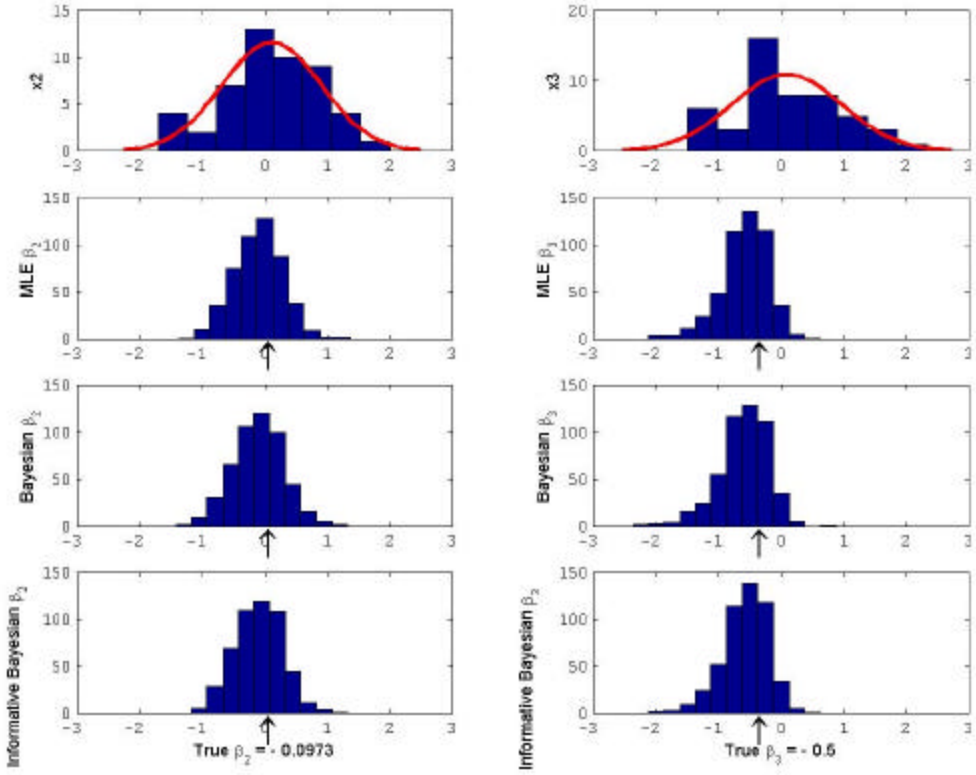


Figure B.3: Sampling Distributions for Design Point 3

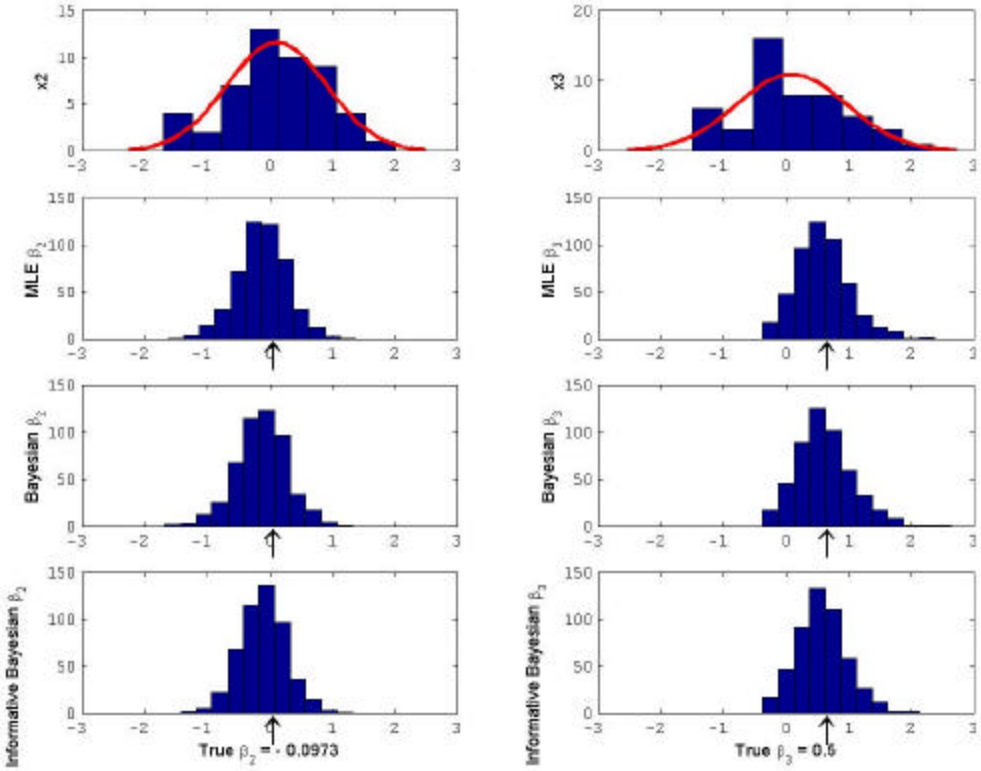


Figure B.4: Sampling Distributions for Design Point 4

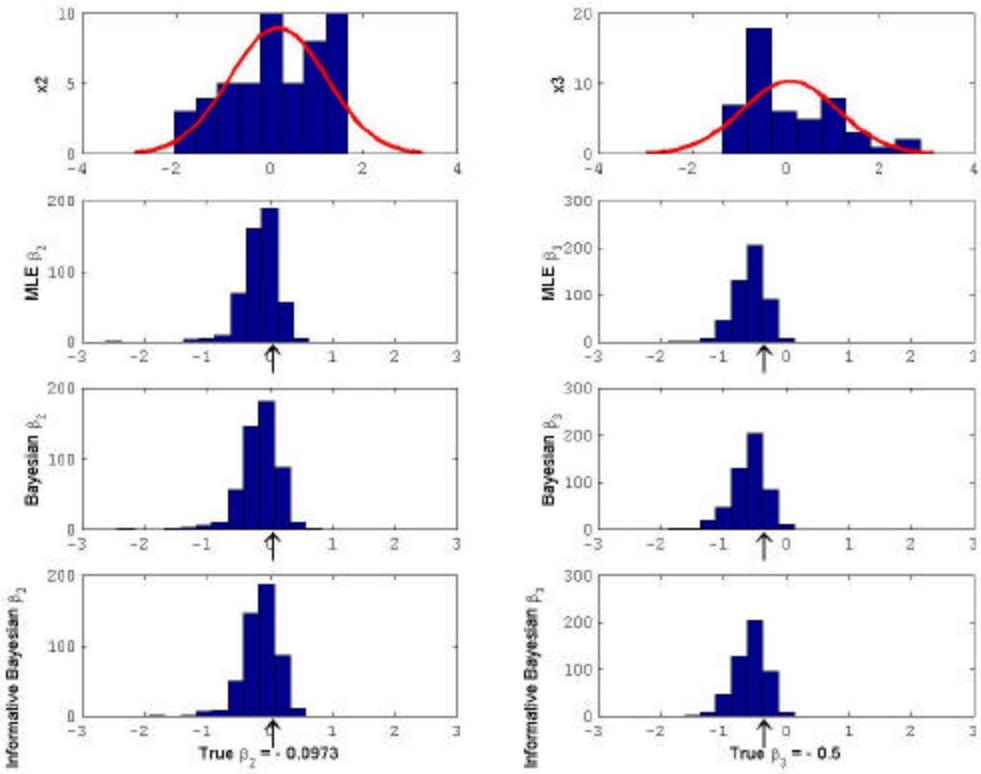


Figure B.5: Sampling Distributions for Design Point 5

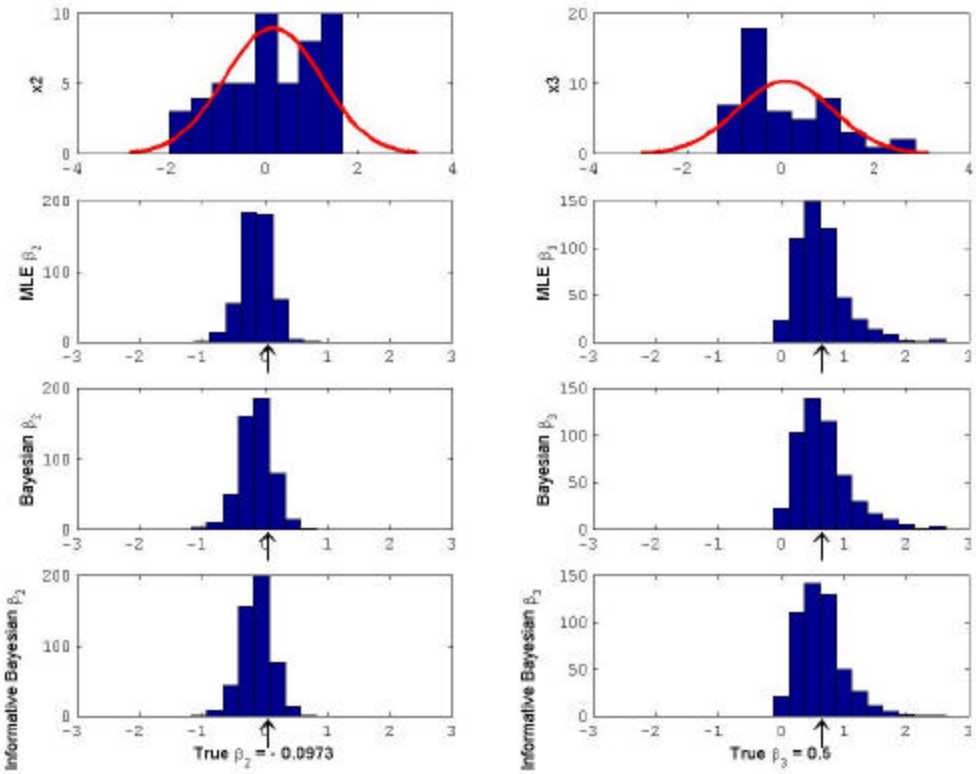


Figure B.6: Sampling Distributions for Design Point 6

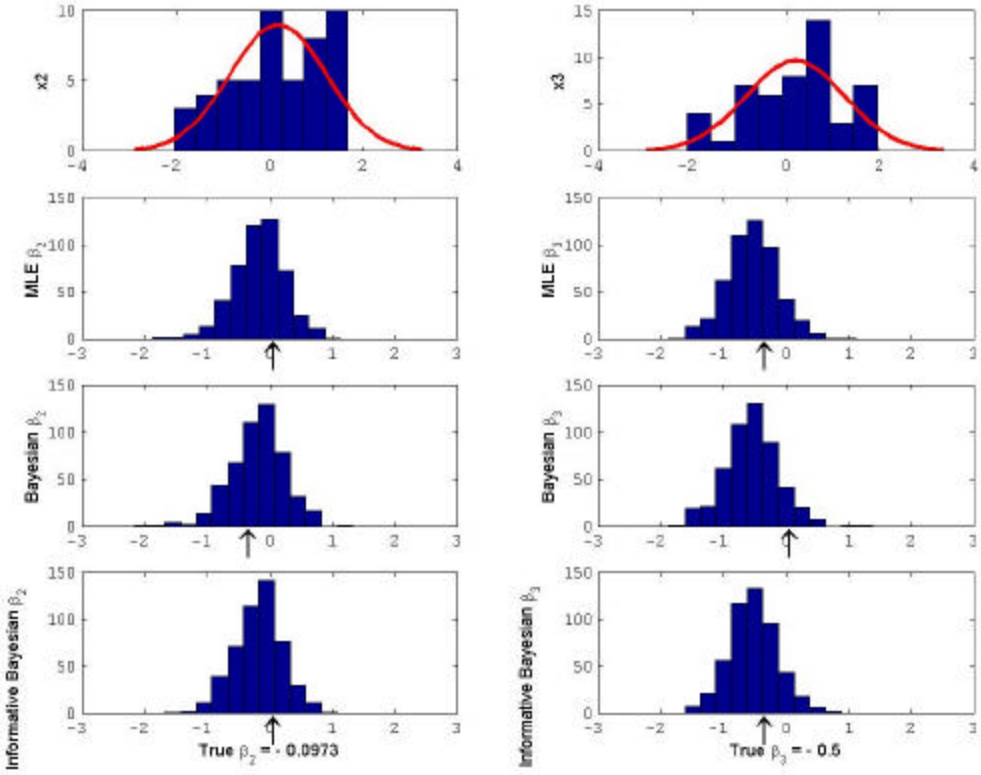


Figure B.7: Sampling Distributions for Design Point 7

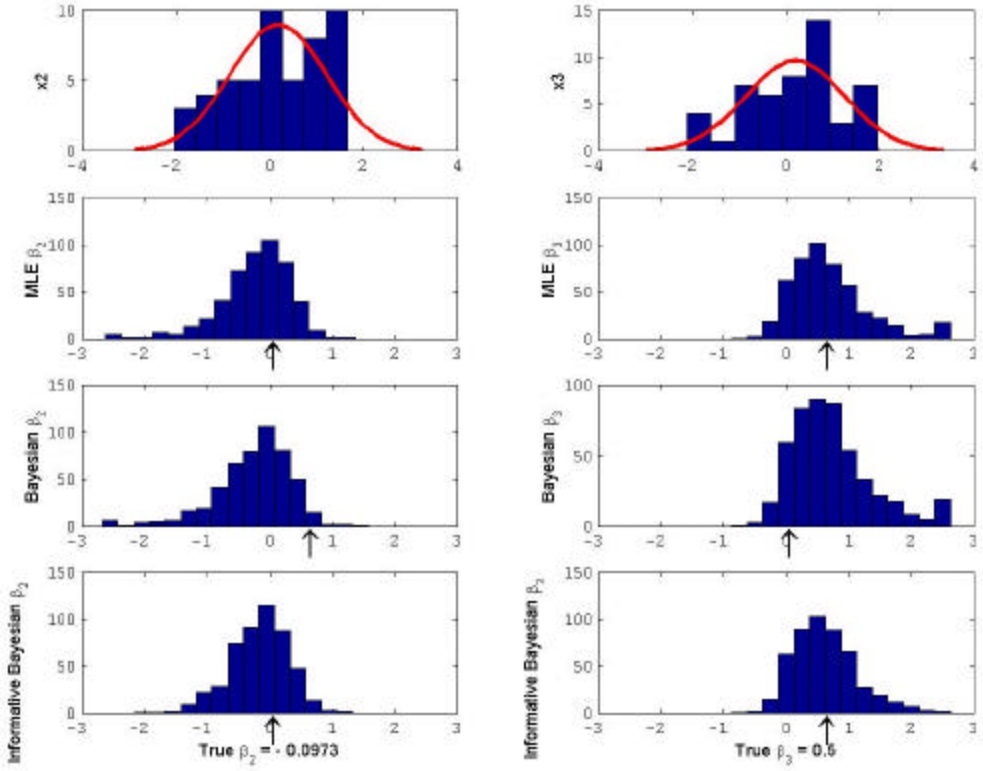


Figure B.8: Sampling Distributions for Design Point 8

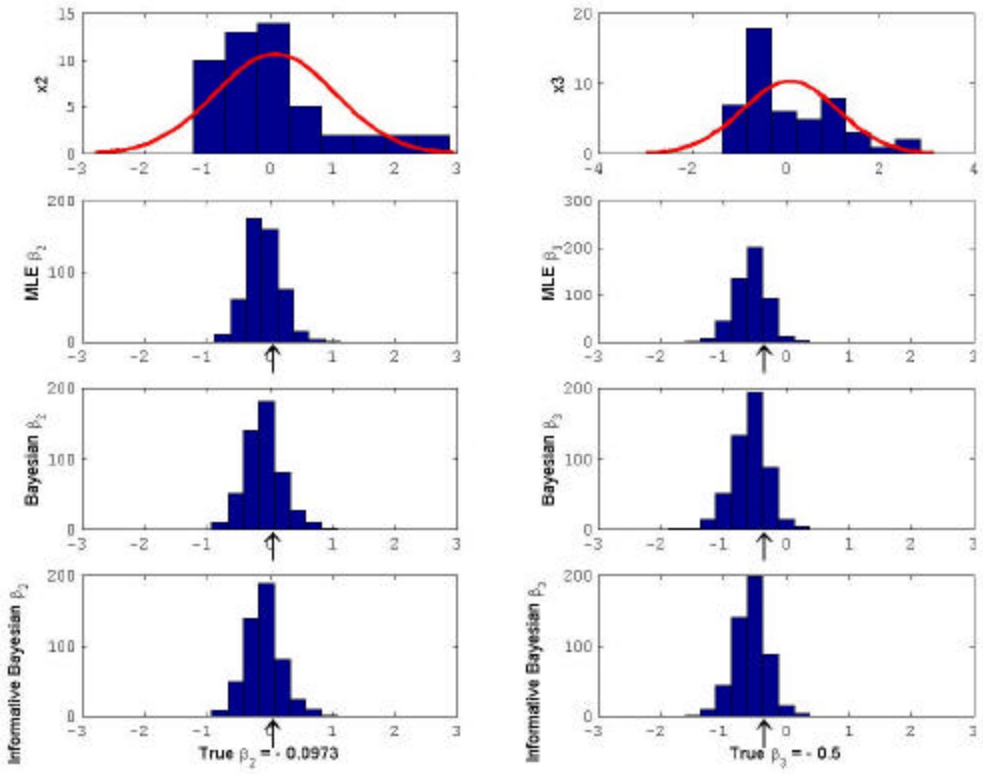


Figure B.9: Sampling Distributions for Design Point 9

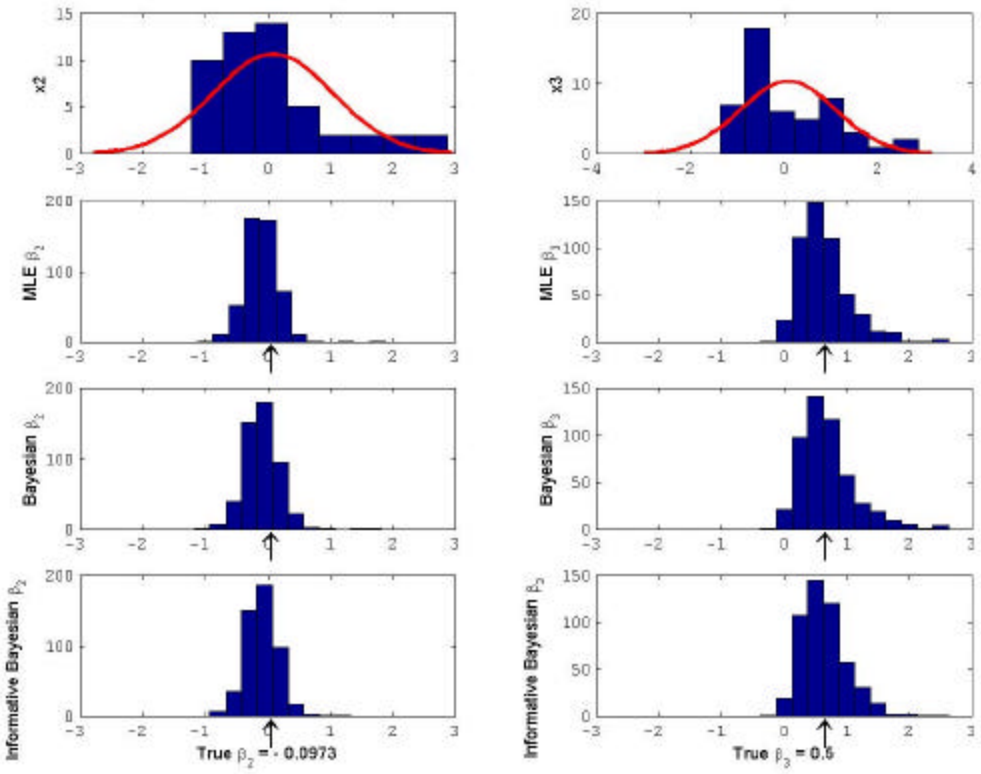


Figure B.10: Sampling Distributions for Design Point 10

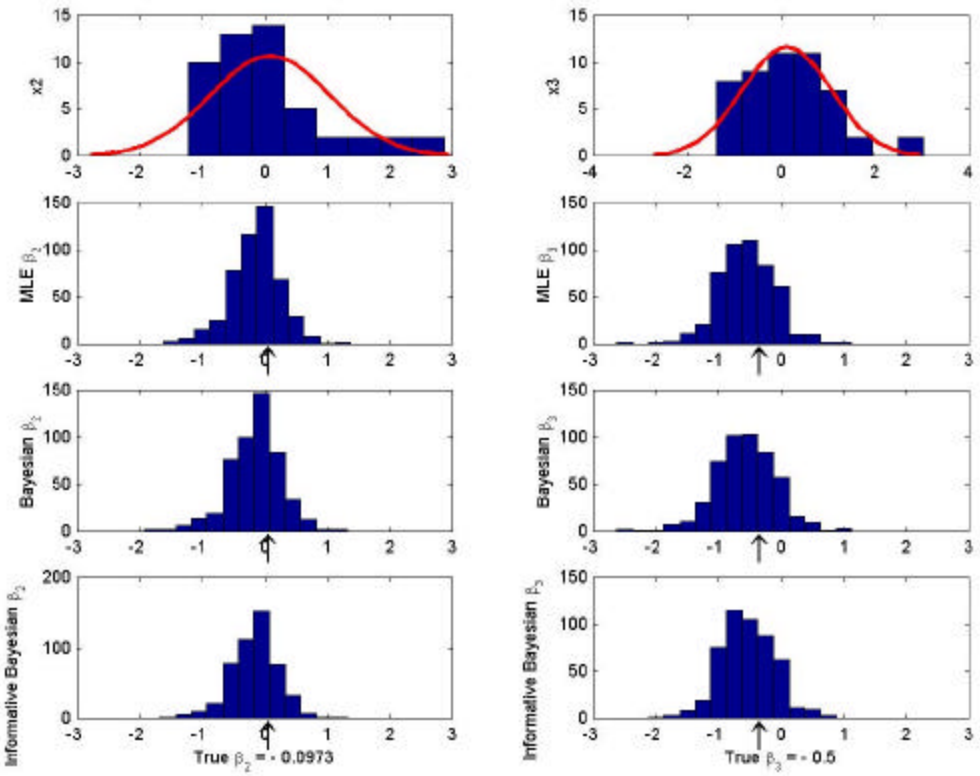


Figure B.11: Sampling Distributions for Design Point 11

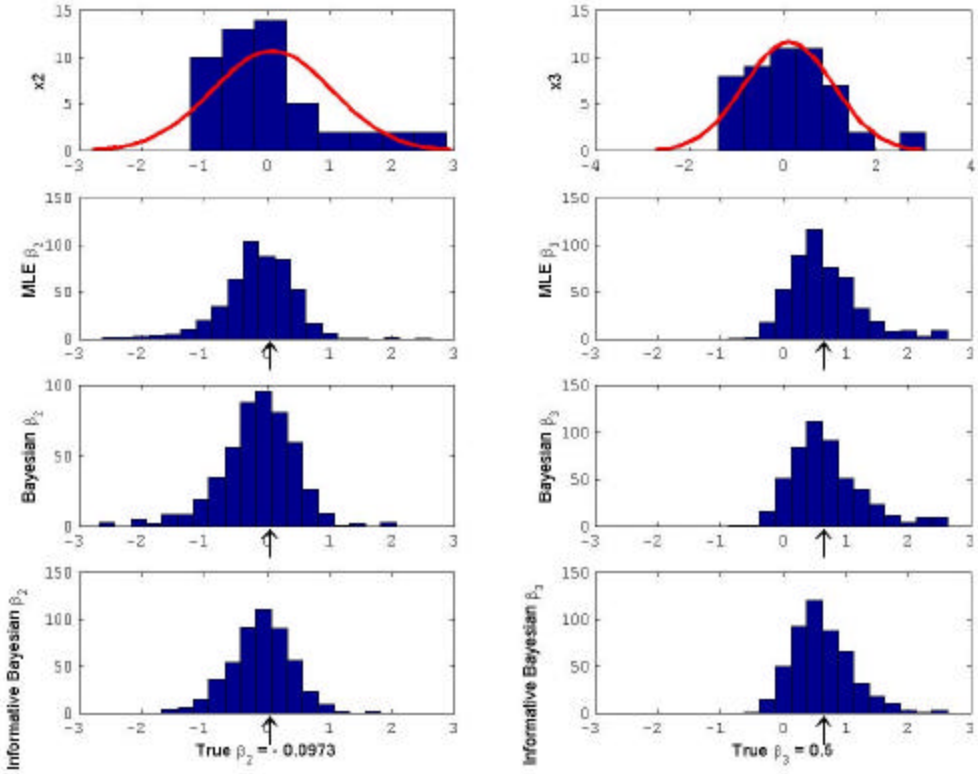


Figure B.12: Sampling Distributions for Design Point 12

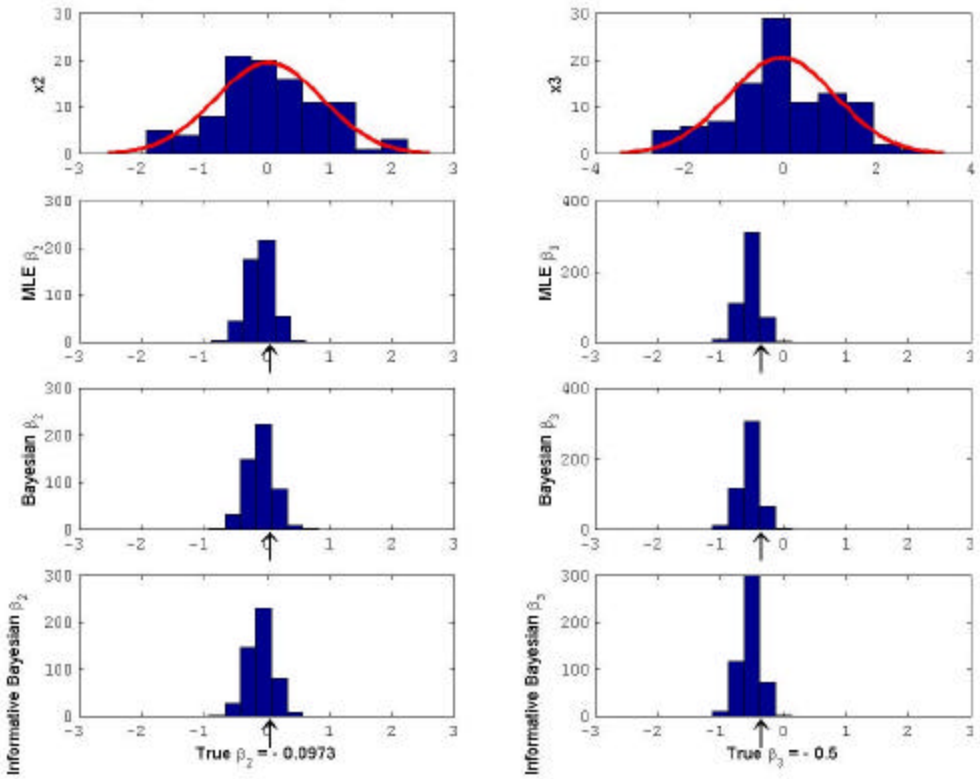


Figure B.13: Sampling Distributions for Design Point 13

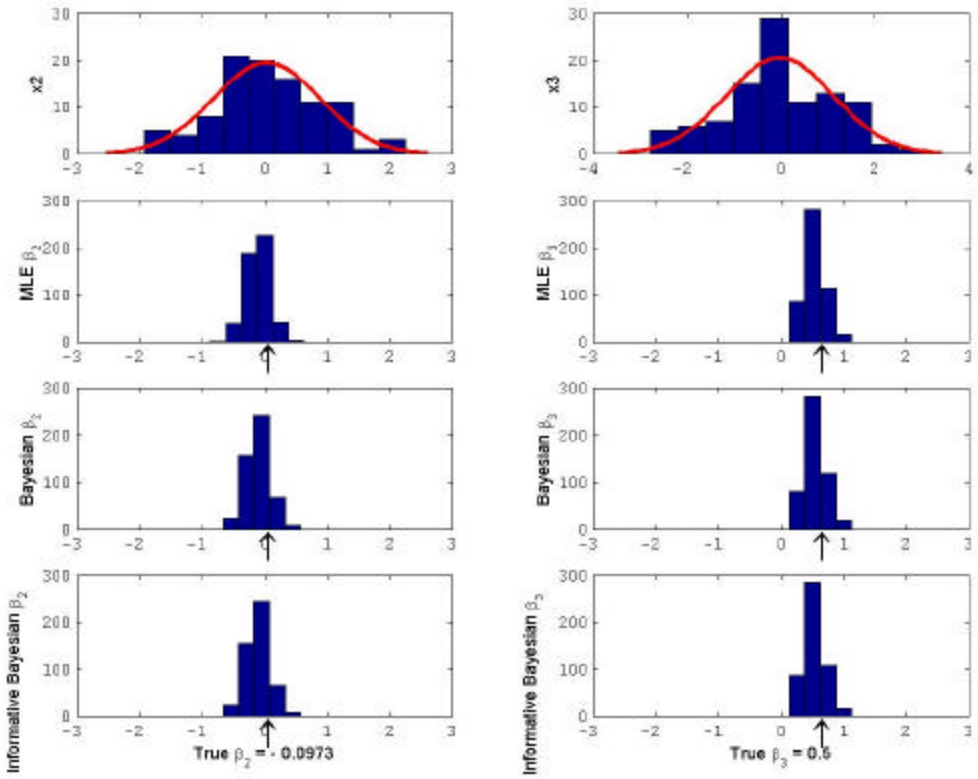


Figure B.14: Sampling Distributions for Design Point 14

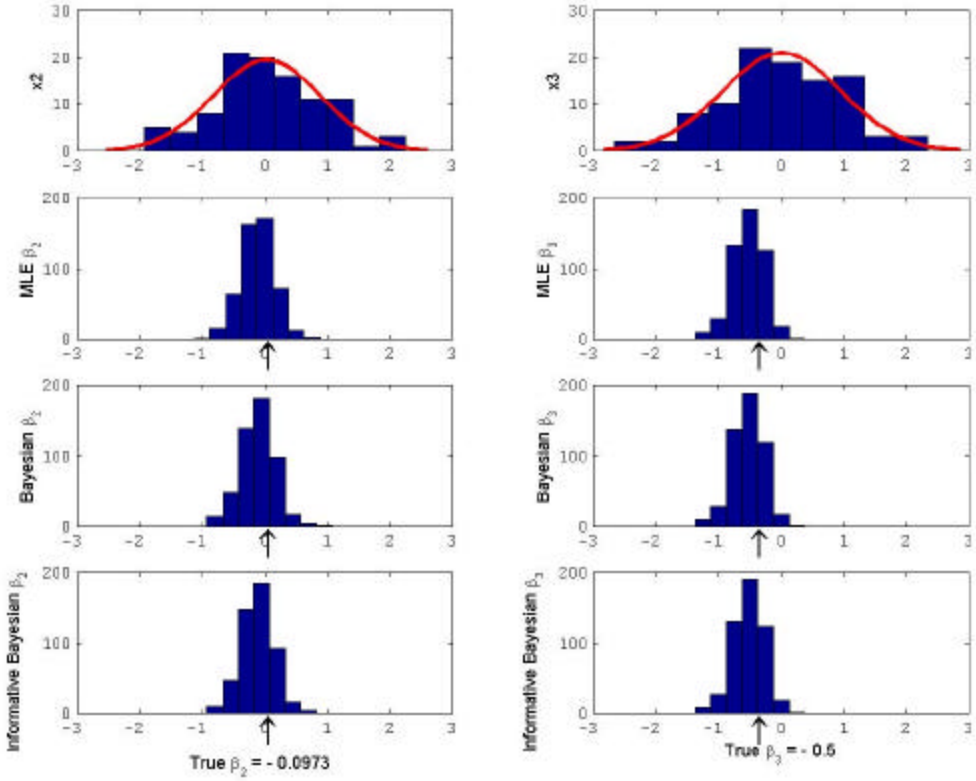


Figure B.15: Sampling Distributions for Design Point 15

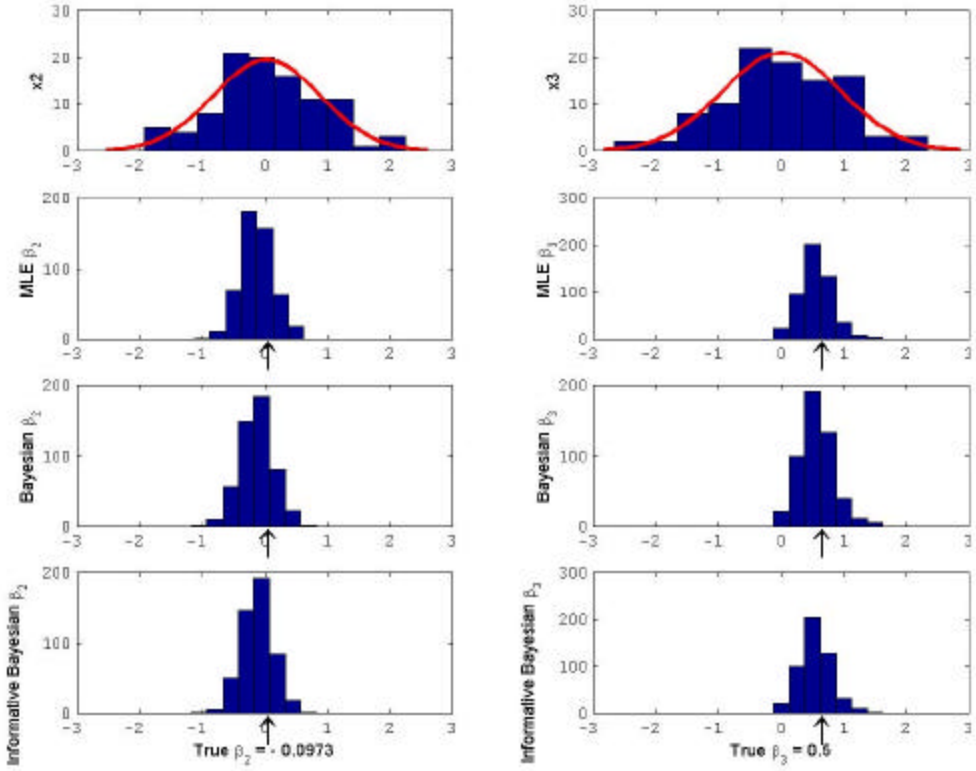


Figure B.16: Sampling Distributions for Design Point 16

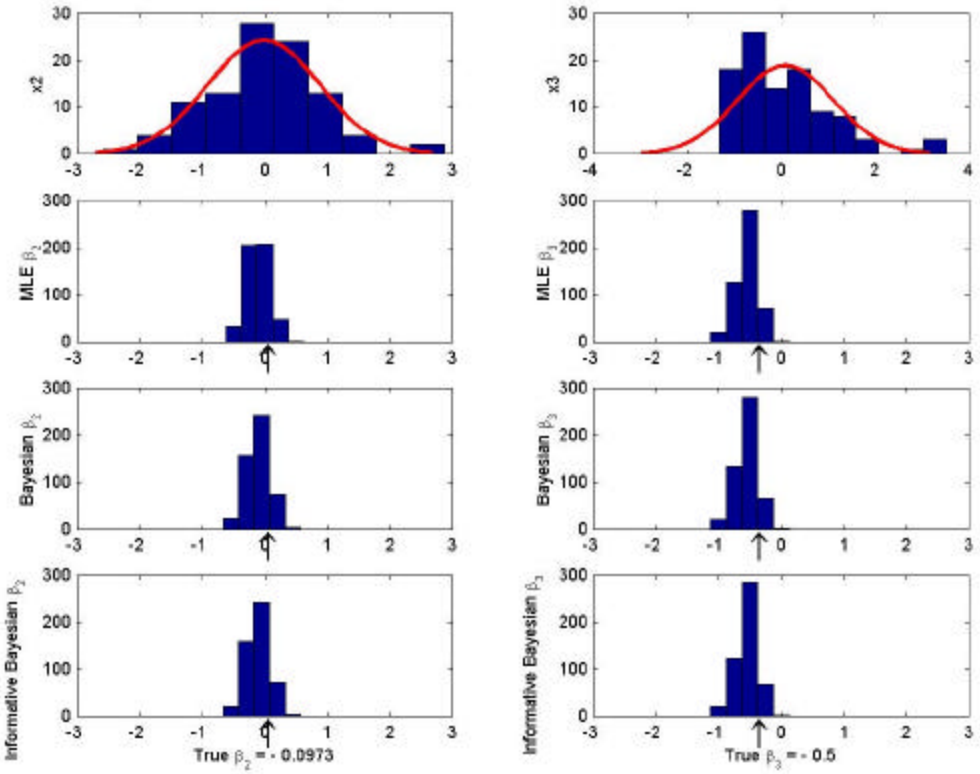


Figure B.17: Sampling Distributions for Design Point 17

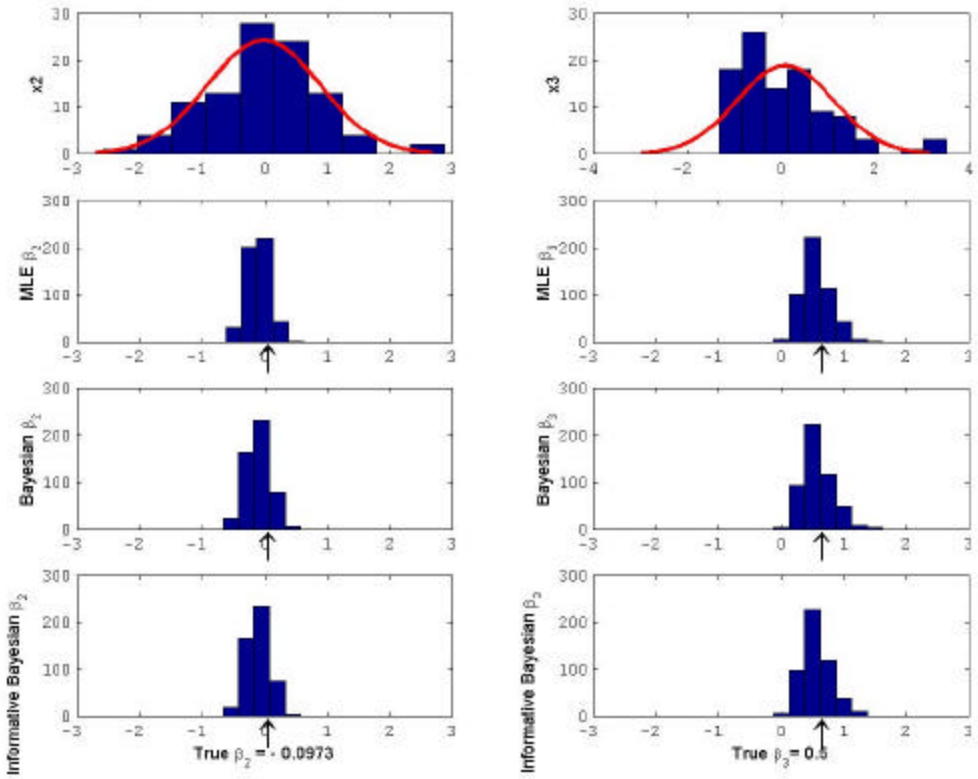


Figure B.18: Sampling Distributions for Design Point 18

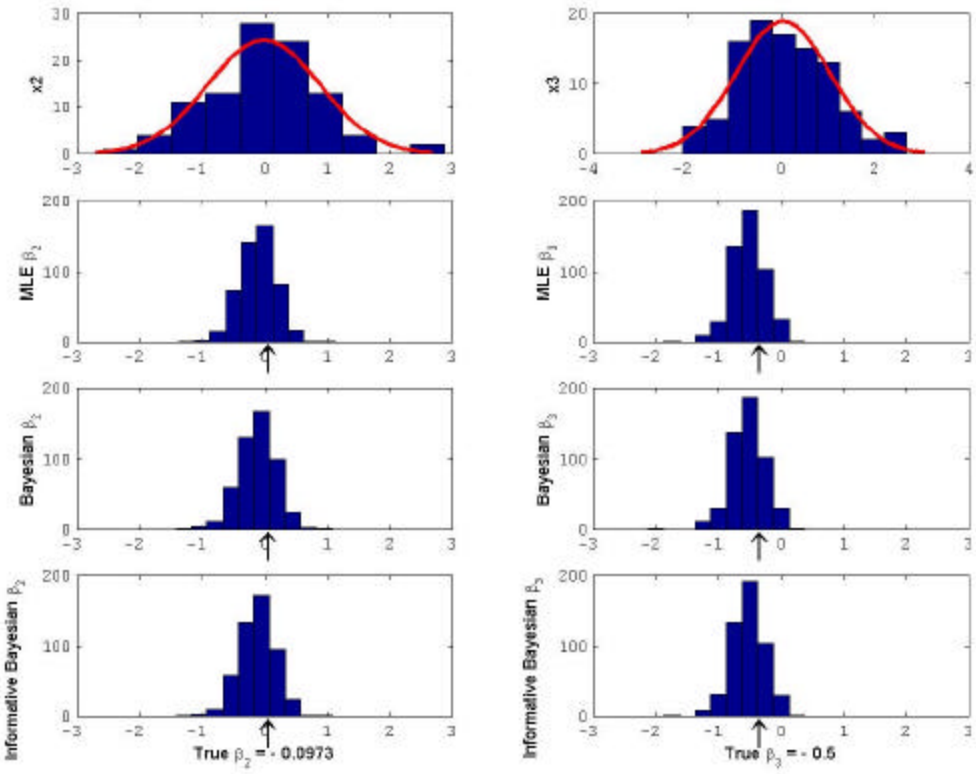


Figure B.19: Sampling Distributions for Design Point 19

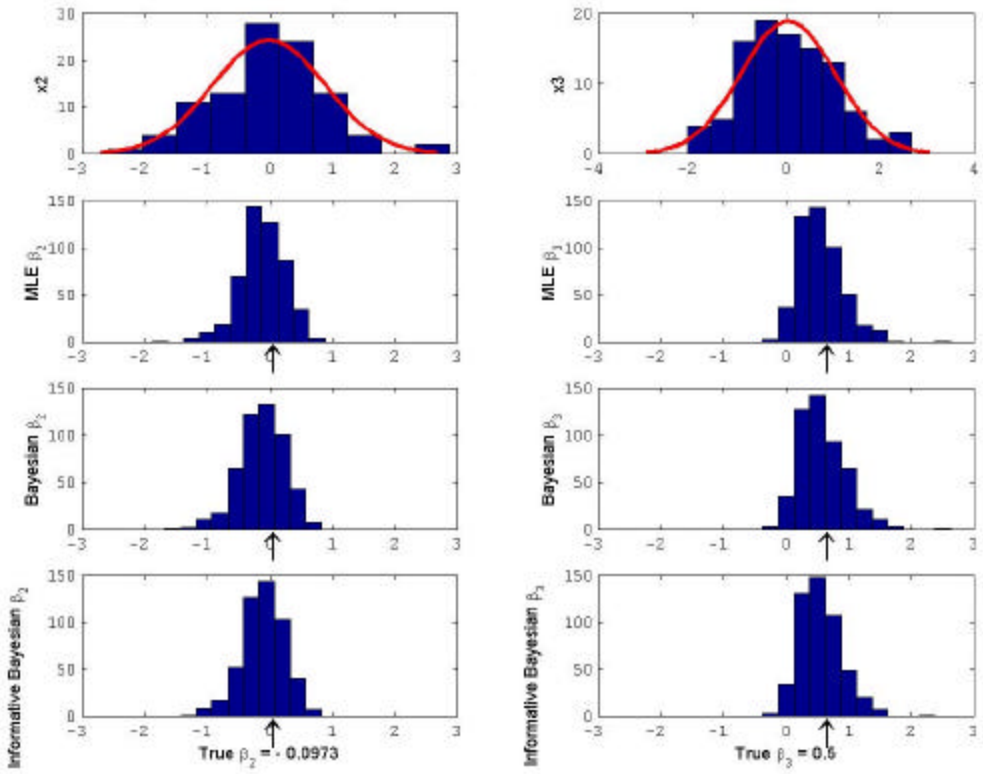


Figure B.20: Sampling Distributions for Design Point 20

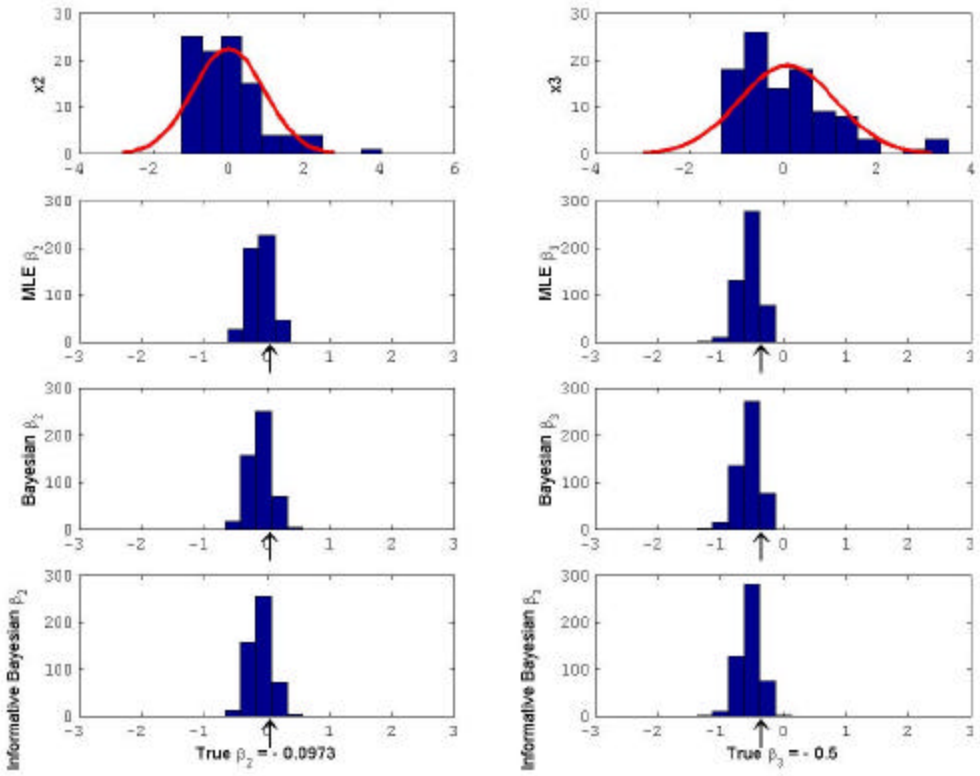


Figure B.21: Sampling Distributions for Design Point 21

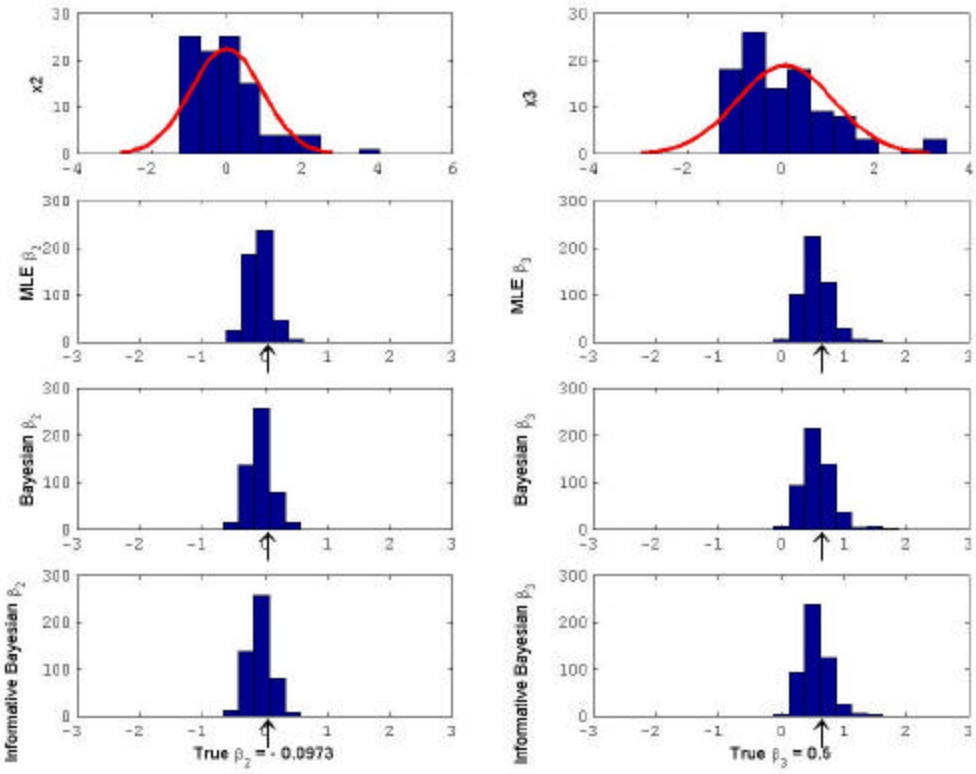


Figure B.22: Sampling Distributions for Design Point 22

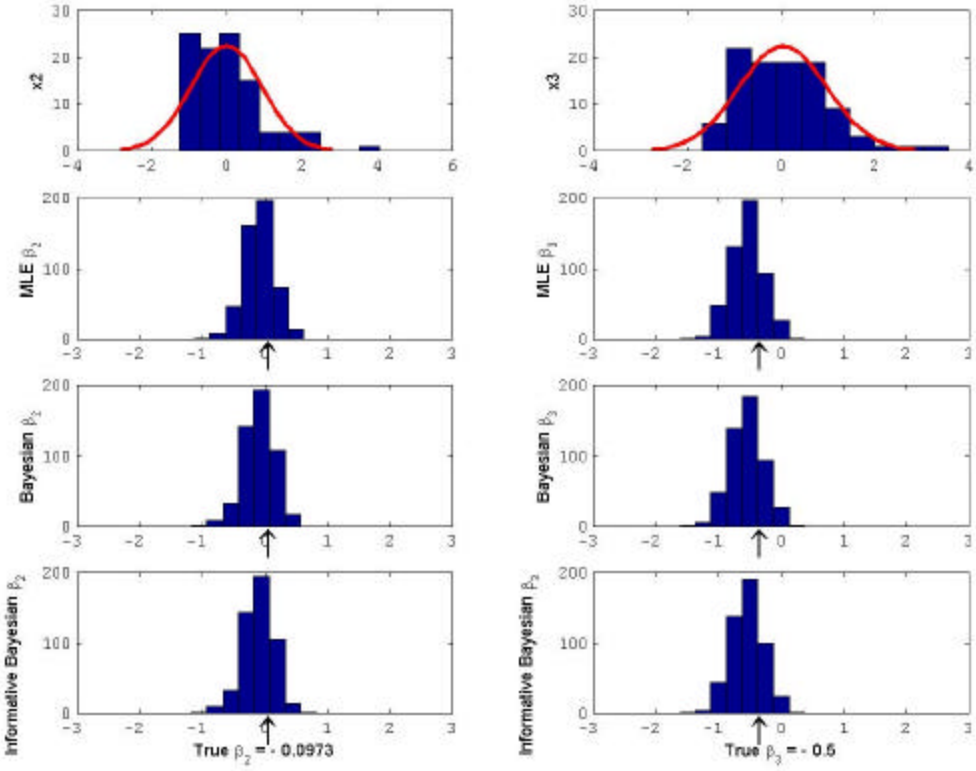


Figure B.23: Sampling Distributions for Design Point 23

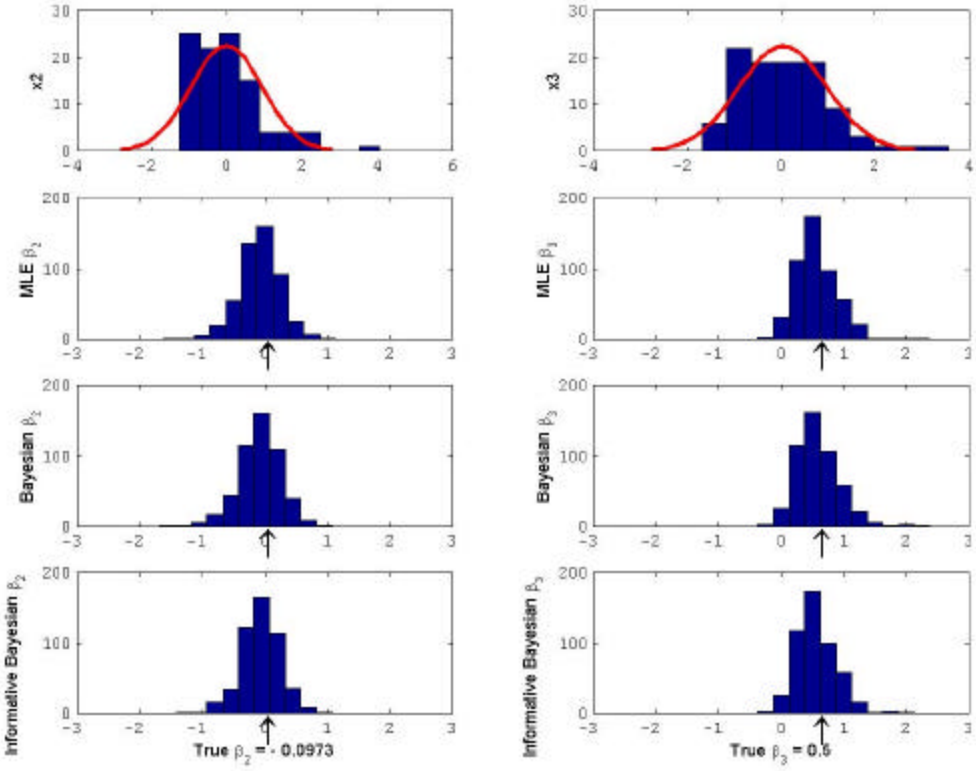
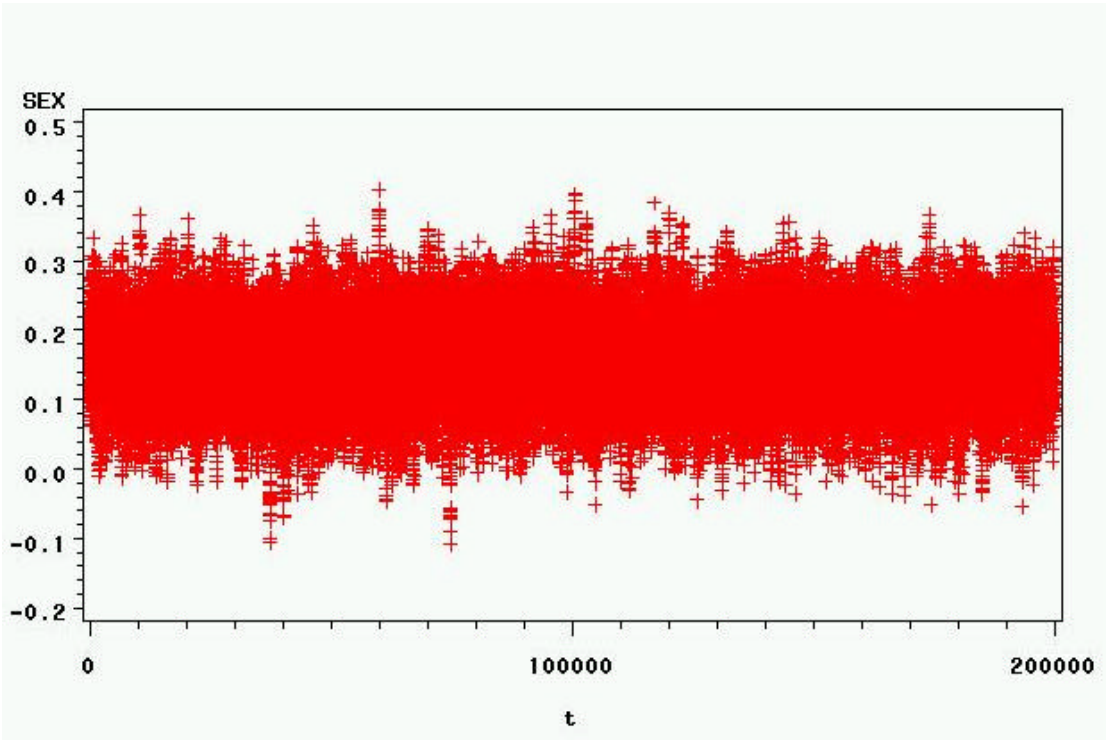


Figure B.24: Sampling Distributions for Design Point 24

APPENDIX C: ADDITIONAL PLOTS FOR CHAPTER 4



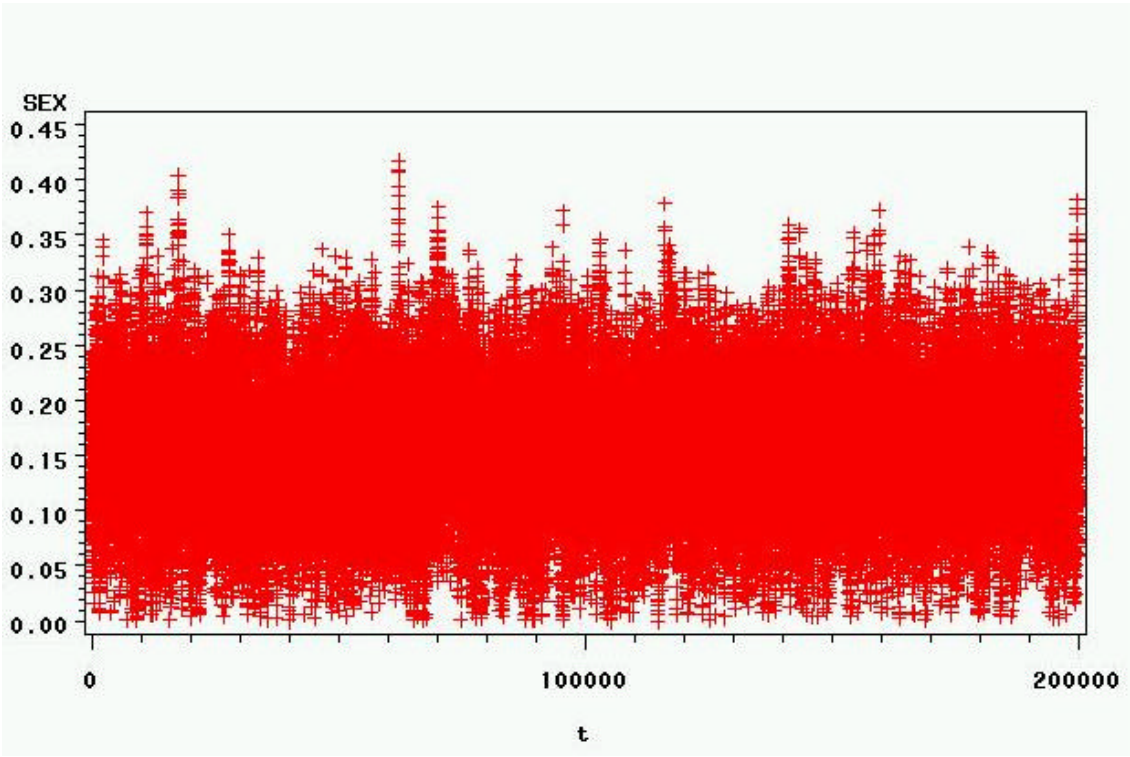


Figure C.1: Coefficient of SEX (Unconstrained and Constrained Models)

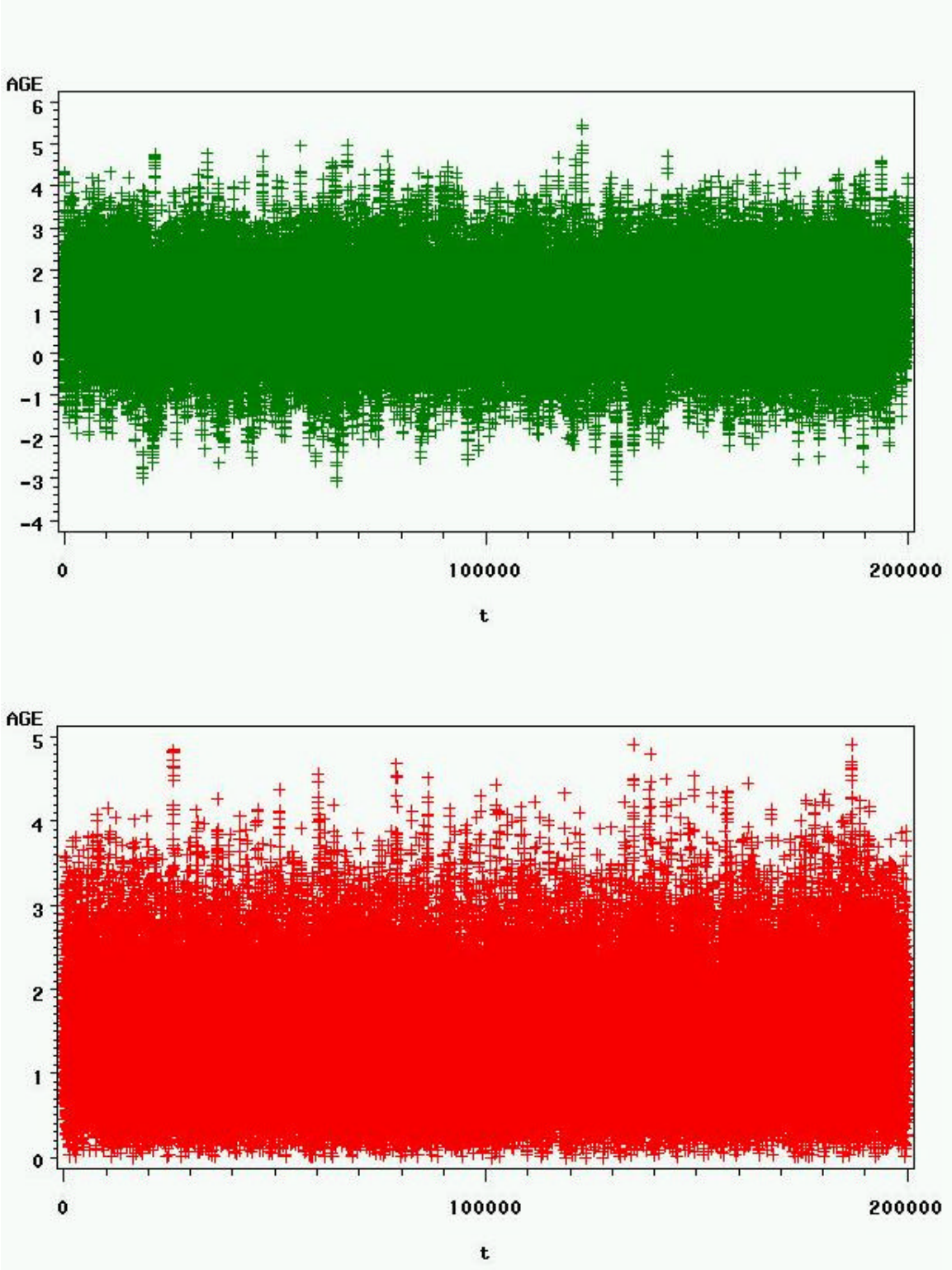


Figure C.2: Coefficient of AGE (Unconstrained and Constrained Models)

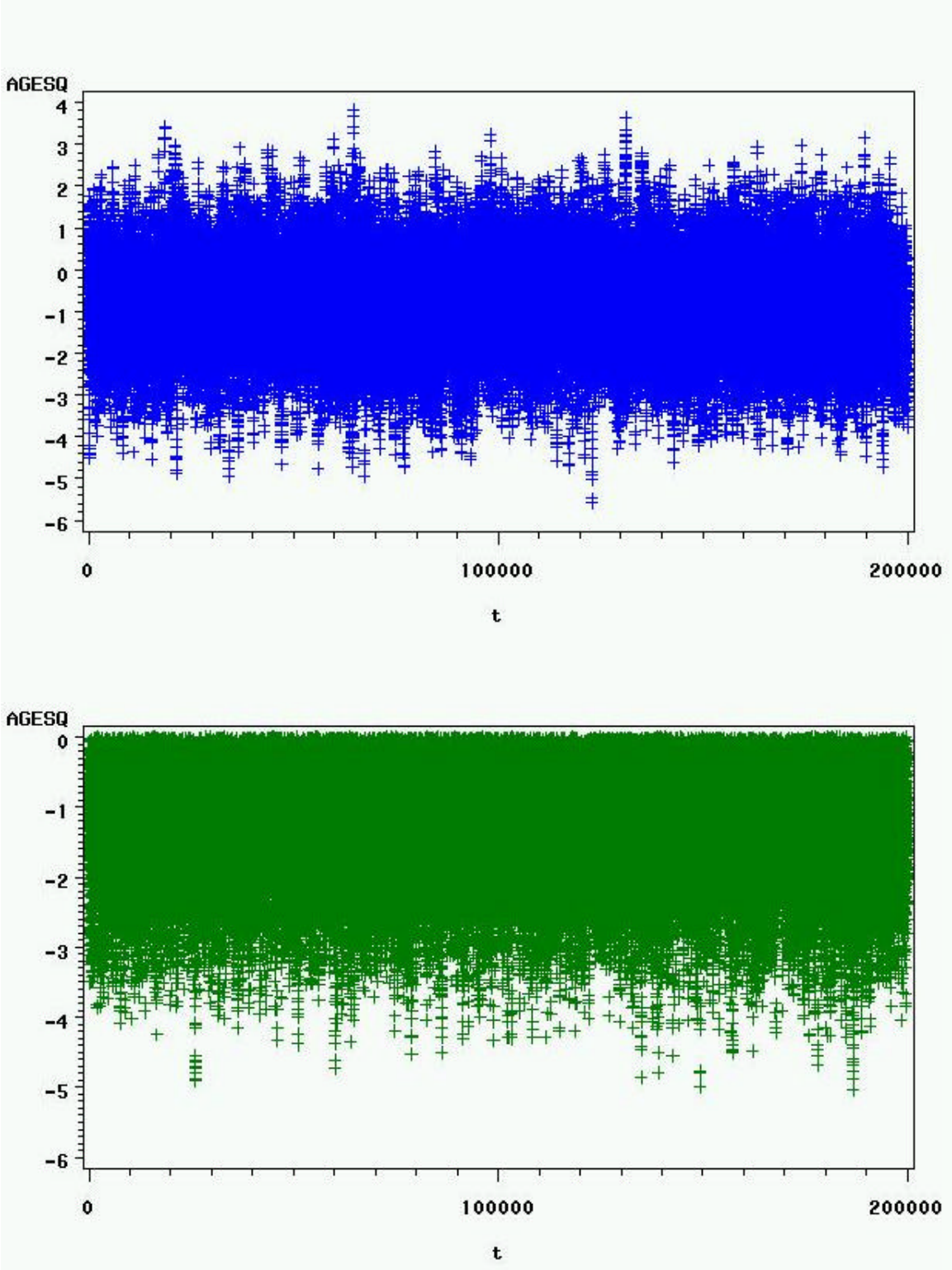


Figure C.3: Coefficient of AGESQ (Unconstrained and Constrained Models)

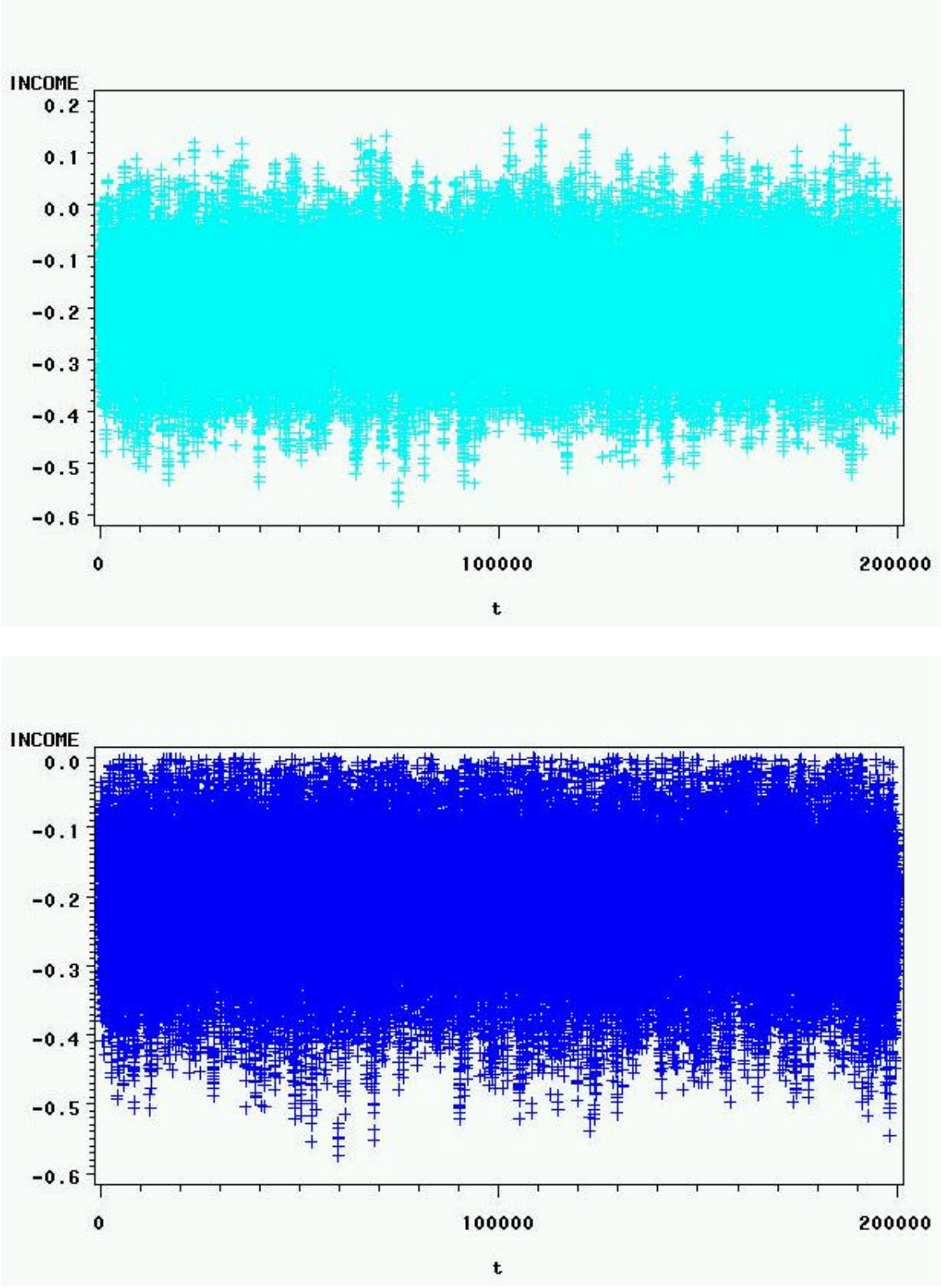


Figure C.4: Coefficient of INCOME (Unconstrained and Constrained Models)

VITA

Asli K. Ogunc was born in Izmit, Turkey, to Nur and Erden Kavaklioglu. She was raised in Izmit, Canakkale, Ankara and Istanbul with her brother, Serkan Kavaklioglu. She received her Bachelor degree from Marmara University in 1990 and got married to Kurtay Ögunc the same year. Asli got her Master of Business Administration degree from Western Michigan University in 1992 and Master of Science in Economics from Louisiana State University in 1996. She taught in the Economics Department of the Louisiana State University for four years. During that time Asli and Kurtay had a baby girl, Patara in July of 1998. She began working as a statistician at Capital One Financial Corporation in August 2000. Asli will complete the degree of Doctor of Philosophy in May 2002.